Biogeographical Analyses and Applications: The Study of Plant Distribution Patterns in West Africa

Dissertation

zur

Erlangung des Doktorgrades (Dr. rer. nat.)

der

Mathematisch–Naturwissenschaftlichen Fakultät

der

Rheinischen Friedrich–Wilhelms–Universität Bonn

vorgelegt von

Jaime Ricardo García Márquez

aus

Bogotá, Kolumbien

Bonn, März 2010

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn.

1. Gutachter: Prof. Dr. Wilhelm Barthlott

2. Gutachter: Prof. Dr. Pierre. L. Ibisch

Tag der Promotion: 15 June 2010

Erscheinungsjahr: 2011

ACKNOWLEDGEMENTS

This thesis was carried out at the Nees Institute for Biodiversity of Plants, Rheinische Friedrich-Wilhelms-Universität Bonn, headed by Prof. Dr. Wilhelm Barthlott, and builds on research on continental and global patterns of plant diversity conducted by the BIOMAPS working group. The study is embedded in the framework of the BIOTA West Africa Project, funded by the German Federal Ministry of Education and Research. It was also supported by the project "Biodiversität im Wandel" funded by the Akademie der Wissenschaften und Literatur, Mainz.

It is a pleasure to thank those who made this thesis possible.

First of all I would like to express my sincere gratitude to my supervisor, Professor Dr. Wilhelm Barthlott, who supported me and trusted me from the first day. He encouraged me and allowed me to participate in different congresses and courses where the knowledge acquired was of primary importance to elaborate my research ideas. All that knowledge was not only useful for making my own research but it will be essential for my future carrier. It was also a pleasure to participate in your lectures which introduced me to a broader perspective of global biogeography.

I would like to show my special gratitude to Dr. Jan Henning Sommer. He provided me with many helpful suggestions, important advice and constant encouragement during the course of this work. But especially for your unconditional, immediate and endless support, personally and professionally, every time I needed it. And I needed it a lot.

I wish to thank Professor Dr. Pierre L. Ibisch (University of Applied Sciences Eberswalde) who kindly accepted to co-supervise this thesis. Your comments and suggestions helped me to improve the quality of the final work.

My special gratitude goes to Dr. Carsten Dormann (UFZ). Thank you for finding the time (which I know is very scarce) to read and listen to my ideas. For your suggestions and your always availability to help me with any issue I needed to solve.

I owe my deepest gratitude to my colleagues Sylvestre Da, Katharina Sabellek and Katja Seis. I could not think of better colleagues to be and to work with. I thank you for sharing your ideas with me, for listening to my ideas and your always helpful suggestions. But above all of that I thank you so much for your friendship and for your care to my family. And forgive me for all those moments I bothered you making you see and hear about my incomprehensible R and GRASS code and graphics and analysis, etc... but who else could I bother?

Many people also contributed in a way or another to the successful completion of this thesis. I am indebted to all of them for their suggestions, their support, constructive critic and for provided me with a nice working atmosphere. They include Dr. Daud Rafiqpoor, Prof. Dr. Holger Kreft, Dr. Wolfgang Küper, Dr. Jens Mutke, Dr. Nils Köster, Pascal Wauer, Sebastian Eschweiler (thank you for showing me the way to Linux), Laurens Geffert, Christian Haase.

Thanks to the numerous scientists and institutions who contributed to the development of the Biogeographic Information System on African Plant Diversity (BISAP) and the West African Plant Database, especially to all those African students who spent a lot of time in the field collecting data.

I would also like to thank Dr. Jakob Fahr and Matthias Herkt from Ulm University and Dr. Johannes Penner from Museum für Naturkunde in Berlin for sharing their data on bats and amphibians respectively and for very fruitful discussion on species modeling.

Durante mi trabajo de doctorado ocurrieron eventos personales que fueron difíciles de sobrellevar. De hecho, no lo hubiera logrado sino hubiera sido por la fortaleza, las energías positivas y el amor que todos ustedes, mi familia, constantemente me brindaron. Gracias a todos por su apoyo, por creer en mi, por su ejemplo de vida y su amor. Circe mamá gracias por tu paciencia y tu esfuerzo con las correcciones del Inglés que sin ellas la tesis no hubiera tenido algún sentido.

My Olga, what a period of time we had together. I thank you so much for all the strength you gave me in many times of weakness. For listen to me and share my ideas and thoughts. You are a wonderful woman, a wonderful mother, a wonderful wife and I love you. And please forgive me for all that time I spent in front of the computer when I should be with you.

My children, you are my motivation, my happiness, my energy, and the reason to do things right. Ricardo my son, Maia my little princess, thank you for bringing so much happiness

and proud to my life. And thank you guys for letting me know that science is not that important after all.

To all my family. You educated me with principles and you taught me to be an honest and an ethic man. I promise you I will keep being that person, personally and professionally. This piece of work is a very small contribution to science, but I did it all by myself. It is the result of my own ideas and my own efforts and it is dedicated to you.

CONTENTS

CHAPTER 1

GENERAL INTRODUCTION

2010 was declared by the United Nations as 'The International Year of Biodiversity'. This declaration is an international recognition of the immense relevance that biodiversity, at all its hierarchical levels (i.e., ecosystems, populations, species, genes), has on human well-beign, ecosystem functions and the services they provide (Chapin et al. 2000, Loreau et al. 2001, Dirzo and Raven 2003, Millennium Ecosystem Assessment 2005, Díaz et al. 2006, Hector and Bagchi 2007). This declaration is also pertinent, and partly originated by the current rates of habitat destruction. Many factors contribute to biodiversity loss, among others, land conversion for agriculture (Tscharntke et al. 2005), urbanization (McKinney 2002), alien species (Sax and Gaines 2003), deforestation , climate change (Root et al. 2005), human growth (McKee et al. 2004).

Despite its importance, there is still no consent on the definition of biodiversity. A quick search in an on-line encyclopedia (i.e., `http://www.encyclo.co.uk/define/BIODIVERSITY`) results in 25 different definitions of the term 'biodiversity'. Because of this ambiguity, there is a large list of different criteria available to measure biodiversity, with species richness (i.e., count of the number of species in a given place at a given time) being the most frequently employed (Currie and Paquin 1987, Gotelli and Colwell 2001, Jiménez et al. 2009). Therefore, species richness has been an omnipresent criterion to select areas for conservation of biodiversity (Scott et al. 1987, Myers et al. 2000).

However, species richness alone is not the most adequate surrogate to define areas for conservation of biodiversity (Fleishman et al. 2006). By protecting areas rich in species, other species that are not located in the same areas might be excluded. For example,

endemic species, whose geographical distribution do not necessarily coincide with areas of high species richness (Orme et al. 2005). Or complementary species, that is, species unrepresented in high species richness selected areas (Williams 2001). Selection of important areas to conserve overall biodiversity must ideally be rich in species and endemism (Myers et al. 2000), and maximize species complementarity (Williams et al. 2006). Other criteria to incorporate into conservation efforts must take into account the changing nature of species richness and the dynamic processes occurring in places of high species richness (Bestelmeyer et al. 2003).

Regardless of which criterion is used to measure biodiversity, the minimal information required is the geographical location describing the distribution of any biological target (e.g. species, ecosystems). Despite the recognition of big gaps of such information in many parts of the world (Balmford et al. 2005), in recent years there has been an increment in the availability of biological databases containing species' geographical records covering local to global scales (Edwards et al. 2000, Graham et al. 2004, Soberón and Peterson 2009). The best example of such a database is the freely accessible 'global biodiversity information facility' (GBIF) database. It serves as a platform for coordination and interoperability of several biological databases located worldwide in order to make them available to any user (Edwards et al. 2000). Biodiversity informatics is a new focus of study and research that seeks to provide the means to share and synthesize all the biological data available and the knowledge it provides (Soberón and Peterson 2004, Guralnick and Hill 2009, Paton 2009).

The final goal of this information is to analyze species distribution patterns across the globe. Furthermore, to understand what are the main factors driving the observed patterns. Knowing where and why are the fundamental prerequisites to objectively built strategies for biodiversity conservation. In that respect, theories and methodologies in biogeography are promising tools to achieve this goal (Whittaker et al. 2005). In fact, Whittaker et al. (2005) proposed 'Conservation Biogeography' as a new field of research and defined it as:

> ...the application of biogeographical principles, theories, and analysis, being those concerned with the distribution dynamics of taxa individually and collectively, to problems concerning the conservation of biodiversity

The current practice of biogeography to tackle conservation issues is partly due to the availability of geographical information systems and remote sensing technologies. These techniques have promoted research on methodologies to properly model and analyze biodiversity distribution patterns, from local to global scales (Luoto et al. 2002, Funk et al. 1999). Available data range from land cover and land use maps, climate layers, digital elevation models, to remote sensing imagery. Parallel to the development and availability

of spatial information, is an increase in the availability of tools to manipulate, process and analyze such information. Noteworthy is the availability of free and open source tools. One of the best examples is the software package SAM (Spatial Analysis in Macroecology) (Rangel et al. 2006), which has been specifically design to perform a range of spatial analysis for macroecology, biogeography and biodiversity conservation. A second group of tools also able to carry out spatial analysis are the set of packages available within the R project for statistical analysis (R Development Core Team 2009). This is an open source program that allows customization of built-in functions for any investigation‘ own goals. The possibilities to carry out spatial analysis are complemented by the availability of other simple and complex statistical tests and approaches and the possibility to create high quality graphics. R is accompanied by a simple and efficient programming language that facilitate automatization of work flows and is currently able to read in and process large size data sets, which is typical for biogeographical applications. All analysis and figures in this thesis were made using the R software.

1.1 Approaches to model species distribution

Ecological niche models play a fundamental role in biogeography. They have been used for a different range of application. For example, analysis of species habitat suitability (Hirzel et al. 2001), analysis of rare species distributions (Le Lay and Guisan 2008), projections of species distribution under land cover and climate change (Thuiller et al. 2006, McClean et al. 2005), invasive species (Thuiller et al. 2005), population viability analysis (Boyce 1992), biodiversity conservation (Rodríguez et al. 2007), among others.

In general, there are two approaches into which niche ecological models can be categorized. The first approach is mechanistic, based on physical relationships between the species and its environment (Yates et al. 2000). The mechanistic approach evaluate species' ecophysiological constraints, life-history, behavioral or genetic plasticity (i.e., intrinsic factors), to predict the species distribution range within its tolerance limits (Kearney and Porter 2009, Robertson et al. 2003). The predictor variables used in mechanistic models tend to be resource or direct rather than indirect gradients (*sensu* Austin (2002)) The second approach is correlative, based on empirical relationships between environmental parameters (i.e., extrinsic factors) and species occurrences (Franklin 1995, Scott et al. 2002, Guisan and Zimmermann 2000). Correlative approaches combine a set of predictor variables (e.g. environmental variables) with biological collection data and through some statistical-fitting technique the relationships between the known locations of species and the set of environmental predictors are defined (Guisan and Zimmermann 2000, Robertson et al. 2003). These relationships are then use for extrapolation and mapping. The main disadvantage of the mechanistic approach is that it requires a profound knowledge of the species physiology and that knowledge is very limited to a few species. The main disadvantage of the

correlative approach is that it makes use of distal environmental predictors (i.e., indirect gradients). Austin (2002) defines proximal and distal as "being to the position of the predictor in the chain of processes that link the predictor to its impact on the plant".

Because of the considerably good availability of species geo-referenced records and environmental parameters covering the globe and at different spatial resolutions, the majority of application of ecological niche models have been with the correlative approach. However, there are many assumptions and uncertainties included in this type of models that should be carefully evaluated if the final goal is to inform conservationist and decision makers (Wiens et al. 2009, Dormann 2007b).

Another important application of ecological niche models is the possibility to aggregate species distribution ranges predictions to estimate several diversity aspects, like species richness, endemism, and beta diversity. This approach to model spatial patterns of biological diversity has been called 'taxon-based' (Barthlott et al. 1999). Sources of information within this approach include species locality information in grid maps (e.g. Humphries et al. 1999) or distribution atlases (e.g. Heywood 1993), and information stored in natural history collections, museums and herbaria (e.g. Küper et al. 2004b). It is important to note that the success of this approach relies on sufficient data availability, which might be the exception for many organisms and geographical regions. A complementary modeling and mapping approach, 'inventory-based' (Barthlott et al. 1999), uses summary information of local and regional floras and checklists. This approach has been primarily use to map vascular plant richness patterns on a global scale (Barthlott et al. 1996, Kier et al. 2005, Mutke and Barthlott 2005, Kreft and Jetz 2007, Kreft et al. 2008)

1.2 Biogeographical applications in Africa

A detailed description of the history of mapping activities of vegetation patterns in Africa is given in Friis (1999). Descriptions of vegetation patterns in Africa date back at least to the vegetation maps exposed in the Berghaus Physical Atlas (Berghaus 1849). The quantitative information presented in the next generation of maps notably improved due to extensive research and intensive floristic inventories done during the second half of the 20th century. The best examples of maps of plant species richness patterns developed during this time are those by Lebrun (1960) and Ozenda (1982), which have already many similarities to more recent vegetation maps (e.g. Barthlott et al. 1996, Kier and Barthlott 2001, Mutke et al. 2001, Barthlott et al. 2003).

Until this point today, the 'inventory-based' approach (as described in section 1.1) has been always used for cartographic efforts. This approach uses coarse-scaled mapping units (e.g., ecoregions) and the total number of species per mapping unit is estimated based on inventories and species-area relationships (Kier et al. 2005). With even more information

available, and the advent of geographical information systems and remote sensing imagery there was a change towards the 'taxon-based approach'. The focus of research advanced from merely diversity patterns description (i.e., species richness patterns) to explore geographical patterns of other biological organizational levels (e.g. phytocoria) (see Linder et al. 2005), studies on collective properties of plant diversity (e.g., species richness and endemism) (Schmidt et al. 2005, Kier et al. 2006, Küper et al. 2006), identification of data-deficient areas and to define priority areas for conservation (Küper et al. 2006).

Considerable work has been conducted in Africa applying ecological niche models to map the distribution of species richness (Schmidt et al. 2005, Küper et al. 2006). However, the method most commonly utilized (i.e. GARP) has been categorized as a poor performance algorithm (Elith et al. 2006). Additionally, the GARP algorithm produces binary predictions with a great spatial variability from one run to the next. To find the best agreement between different runs a best-subset procedure has been implemented by different authors (Anderson et al. 2003, Raxworthy et al. 2004).

Biogeographical analysis at global and continental scales agree upon the rainforest in West Africa (also known as the Upper Guinean forest) as a center of high diversity (Myers et al. 2000, Küper et al. 2004a, Mittermeier et al. 2005). Biodiversity studies in West Africa are therefore mostly concentrated in these areas, specifically, on floristic and distribution descriptions of woody plants species (Poorter et al. 2004a, Hawthorne and Jongkind 2006). As in any other highly diverse region of the world, forested areas in West Africa have been drastically reduced by many factors, being important logging and agriculture. Further regions in West Africa have also been the focus of research in recent years. A good example are biogeographical applications in Burkina Faso (Schmidt et al. 2008; 2005, Thiombiano et al. 2006). These studies explore species richness gradients and life forms distribution patterns based on field observations and specimen data collected from herbaria. Given the high percentage of forest areas in Ivory Coast and the past and current threats imposed to them, description of the vegetation of this country has received much attention (Ake Assi 2001; 2002, Gautier et al. 1999, Chatelain et al. 2001).

1.3 Institutional Background

This thesis has been conducted at the Nees Institute for Biodiversity of Plants (`www.nees.uni-bonn.de`), at the Rheinische Friedrich-Wilhelms Universität Bonn. One of the institute's scientific research focus has been on macroecological analysis of broad-scale patterns and mapping of biodiversity (Barthlott et al. 1996; 1999, Mutke et al. 2001, Barthlott et al. 2003; 2005, Kier et al. 2005, Mutke and Barthlott 2005, Barthlott et al. 2007, Kreft and Jetz 2007, Kreft et al. 2008). This task has been carried out by the BIOMAPS working group.

The BIOMAPS working group is part of the BIOTA Africa Project "Biodiversity Monitoring Transect Analysis in Africa" (`www.biota-africa.org`) since its beginning in 2001. This project is a cooperative and interdisciplinary research project initiated and funded by the German Federal Ministry of Education and Research (BMBF). It involves a number of German and African institutions and scientists working together towards the sustainable use and conservation of biodiversity in Africa. The BIOMAPS working group within the BIOTA project was part of the sub-project entitled "Analysis of the African Biodiversity and development of sustainable conservation strategies integrating the effects of climate change and land-use". Its aims were to understand the mechanisms responsible for patterns of plant diversity using Africa as a model continent (Kier and Barthlott 2001, Mutke et al. 2001, Küper et al. 2004a;b, Küper et al. 2006) and to understand how plant diversity might change under climate and land use change (McClean et al. 2005; 2006). The second geographical focus of this sub-project was in West Africa ,within the BIOTA-west regional network task. This thesis is embedded within the aims and goals of this task, namely to identify the drivers and processes leading to biodiversity loss, developing methods for the preservation of biodiversity at various spatial scales integrating scenarios on the effects of global change and creating and proposing tools that contribute to the sustainable use of biodiversity

1.4 Aims of this study

Within the 9 years of the BIOTA Africa project, databases containing geo-referenced records of vascular plants and other organisms have been established. These can be considered as the most comprehensive databases currently available for the region. They provide the source of information required to carry out biogeographical analysis. For example, to estimate species richness and endemism patterns, evaluate species richness congruence, calculate the influence of species distribution under climate change, among others. The final purpose of these analyzes is to guide conservation strategies and resource management.

In the first part of this thesis, databases of vascular plants in Ivory Coast, Burkina Faso and Benin were used. The information in the final joint database has been filtered to select all records with a minimal spatial resolution of 10 km^2

In the first study, this database was employed as a case example to develop a methodological framework to quantitatively evaluate the spatial quality of biological databases. A final milestone result is a cartographic representation of an index (i.e., 'gap selection index') describing areas that have been well investigated and areas where more information is needed. A series of spatial analyzes were run to answer several question related to the spatial quality of the database:

- How is the spatial configuration of collection localities in the database? Do they follow a random or a clustered distribution?
- How is the density of collection localities distributed in the study area?
- Is the distribution of collection localities bias towards more accessible areas (e.g., close to cities, to rivers, to streets)?
- Does the distribution of the collection localities properly represent the variation of environmental conditions in the study area?
- Is the database floristically complete?

The second study makes use of the same database to test different modeling approaches to deal with spatially biased information. The maximum entropy technique was used to model the spatial distribution of all species in the database. This technique requires information of the environmental conditions where the species occur and of the environmental conditions that characterize the study area (i.e., background data). Three background data sets were prepared for each species. Random background, which is the commonly used approach (Elith et al. 2006). And two background sets meant to deal with bias in the location of the occurrence records: target background, as explained by Phillips et al. (2009) and index background, created selecting random locations but weighted as a function of the 'gap selection index' developed in the previous study. The main questions to investigate were:

- Does model performance improve when dealing with bias in collection records?
- Are model predictions significantly different between the different background treatments?
- What is the influence of biased and non-biased predictions on spatial patterns of species richness?

In the second part of this thesis, databases of vascular plants, amphibians and bats in West Africa were used. The information in these databases is available at a half degree resolution. The third study employs these three databases to investigate geographical patterns of congruence of species richness and endemism richness between pair-wise comparisons of vascular plants, amphibians, and bats. Specific questions to investigate were:

- How are vascular plants, amphibians, and bats richness and endemism patterns geographically distributed in the study area?
- Are geographical patterns of species richness and endemism of the three groups similar?
- Do exist small extent congruence variation between species richness and endemism patterns?
- Do areas of high endemism (i.e. 'hotspots') for the three groups overlap?
- How good is the coverage of natural vegetation and the network of protected areas in 'hotspots' areas?

CHAPTER 2

A METHODOLOGICAL FRAMEWORK TO QUANTIFY THE SPATIAL QUALITY OF BIOLOGICAL DATABASES

Doubt may be an unpleasant condition
- but certainty is absurd

VOLTAIRE

2.1 Abstract

The basic unit for biogeographical analysis is the geographical information contained in biological databases. A database of vascular plants has been assembled for West Africa (i.e., Ivory Coast, Burkina Faso and Benin) containing more than 53,205 georeferenced observation distributed over 2,931 collection localities. The quality of biogeographical applications is positively correlated with the quality of the spatial information contained in the database. Therefore, a very first step must concern the evaluation of its spatial quality. We propose a methodology where a series of spatial analysis are carried out to quantify the quality of the database. Analysis were done in terms of the spatial configuration of the

collection localities, their spatial and environmental bias and inventory completeness. The spatial configuration of the database followed a highly clustered pattern, with a density average of 0.4 collection localities per 10 km^2 but with few areas having more than 100 collection localities per 10 km^2. The distribution of the collection localities is strongly biased with respect to the distance to cities, to the coast, to rivers, to roads and to protected areas. However, the magnitude and rank of the bias factors varied between countries. The same biased pattern was found in relation to different environmental factors where some areas with particular environmental conditions are underrepresented in the database. Inventory completeness was determined by estimating the potential total number of species as calculated by two non-parametric estimates (i.e., first order Jackknife and Bootstrap). When analyzing the database at a 10 km^2 grid cell size, only 13.8% of the cells contained information. From that percentage, 40% of the cells have a complete species inventory. The percentage of complete cells increases as the resolution of analysis decreases from 10 to 120 km^2. Results were integrated into a new index (i.e., gap selection index) that will serve as a guide for future field work campaigns and as a criteria to be aware of the uncertainties related to biogeographical application based on the current database.

2.2 Introduction

Biogeographical studies aim at understanding how living organisms are spatially distributed, which environmental and biotic parameters influence their distribution and how it changes over time (Brown and Lomolino 1998). The main source for those studies is the information contained in biological databases, specifically, lists of species names and their georeferenced location. Therefore, the success of biogeographical applications heavily depends on the quality of the information on biological databases. Based on this information spatial biodiversity patterns from local to global scales can be investigated (Brown and Maurer 1989). One of the main characteristics of biological information is that it is the result of the combination of different data sources (e.g. Küper et al. 2006). Typical information sources are inventories, herbarium and museum collections, atlases, and multiple field-based relevés, among others (Zaniewski et al. 2002).

It has been continuously mentioned and demonstrated how the spatial information contained in biological databases exhibits different degrees of spatial bias (Whittaker et al. 2005), for example towards location accessibility (Nelson et al. 1990) or conservation areas (Reddy and Dávalos 2003), and even to the distance from the place of residence of biologists (Freitag et al. 1998). One consequence of employing biased data to model the distribution of species or communities might be the erroneous description of real distribution patterns. Instead, the distribution and patterns of sampling effort and/or collection intensity is being represented (Williams et al. 2002). Given this constraint, analysis of the nature and amount of bias in a biological database should be an obligated and a first

step towards the evaluation of the quality of biological database information (Romo et al. 2006).

Consequently, a second step should be the evaluation of the database in terms of its completeness in order to understand how representative the database is in characterizing specific property or aspect of biodiversity. The property most widely employed to describe the diversity of an area has been the number of species in it, that is, species richness (Whittaker et al. 2001). It is also one of the main criteria to define important areas for conservation (Myers et al. 2000). Hence, decisions may also be biased when based on incomplete information.

In fact, it has been shown that the total number of species observed is always less than the true number of species, and hence a negative bias estimator (Walther and Moore 2005, their fig.2). For example, Palmer (1990) argued that there will always be species present in a sample plot that are not present in the sampled subplots. This might be especially the case for biogeographical studies at regional, and even at local scales, where a complete sampling scheme covering the whole study area is an utopia.

Several different methods exist to estimate the total number of species in a certain area based on a restricted number of samples. Among them, non-parametric techniques (e.g. Chao, Jackknife 1 and 2, Bootstrap) have been widely used and have constantly outperformed other techniques, such as species-accumulation curves (e.g. Walther and Martin 2001). By comparing the observed against the estimated number of species different indices can be calculated to describe the completeness and representativeness of biodiversity information (Soberón et al. 2007; 2000, Soria-Auza and Kessler 2008). One common approach to investigate the completeness of biological information is to stratify the area based on grouping or pooling factors and then to examine database information completeness in each of them. For example, Parnell et al. (2003) used vegetation classes, forest and non-forest areas, country political divisions and grid cells to identify which areas had received most research effort and therefore, possessed a more complete biological inventory.

Guidelines for decision making concerning the conservation of plant diversity and land use management are normally originated from analysis of species distributions and ecosystems health at local scales (Colwell and Coddington 1994). But the scale at which complete information is available generally contrasts with this need. As an example, Soberón et al. (2007), in their scale comparative study, found an increase in the percentage of areas with no information available with a decrease in the scale of analysis. Multi-scale analysis therefore help to identify the scale at which the data is best suited for analysis of biodiversity.

One of the goals of analyzing bias, completeness and the effect of spatial scale on biological databases is helping to answer the questions if the available information on biological

databases is enough for the biogeographical research questions at hand or how much additional effort still needs to be invested. It also helps on the identification of gap areas that require further research and sampling effort.

In the last nine years, researchers from different institutions and countries have been gathering a biological database consisting of georeferenced locations of vascular plants. The aim of this study is to quantify the quality of this biological database in terms of 1. the spatial bias in the distribution of the collection localities, 2. the causes or origins of bias in the location of collection localities and 3. the floristic completeness of the database and how it varies at different scales. A final milestone result will be the integration of the analysis mentioned above into a gap selection index (GSI) that serves as an identification tool of areas missing information and where additional sampling will improve spatial coverage of the database, environmental representativeness and floristic completeness.

2.3 Material and Methods

2.3.1 Study Area

The study area encompasses 730600 km^2 in the countries included as part of the BIOTA project transect in West Africa (i.e. Ivory Coast, Burkina Faso and Benin) (Figure 3.1). The terrain is generally flat with a mean elevation of 277 meters above see level. However, some mountainous areas at the west side of Ivory Coast reach an altitude of 1500 meters above see level.

The study area is characterized by a continuous climatic North/South-west gradient. Annual mean temperature ranges from 29.6 degrees Celsius in the northern part of the study area in the Sahelian region to around 18.8 degrees Celsius in the western part of Ivory Coast. Total annual precipitation shows the opposite gradient: it ranges from 300 mm per year in the northern regions to more than 2,600 mm per year in the south–west.

2.3.2 Vascular Plant Species Database

The database used in this study is the result of the compilation of several different heterogeneous sources. Data for Burkina Faso has been described in Schmidt et al. (2005). The database covering Ivory Coast was originated in the botanical garden of Geneva as part of the GIS-Ivory project (Chatelain et al. 2001). These databases were filtered to select the records with a minimal spatial accuracy of 10 km^2.

The final database consists of a total of 53205 observations which are distributed over 2931 collection localities (Figure 3.1, Table 2.1). Collection localities are here relevés, georeferenced herbarium collections, or points extracted from atlases and gazetteers. Only 13.8%

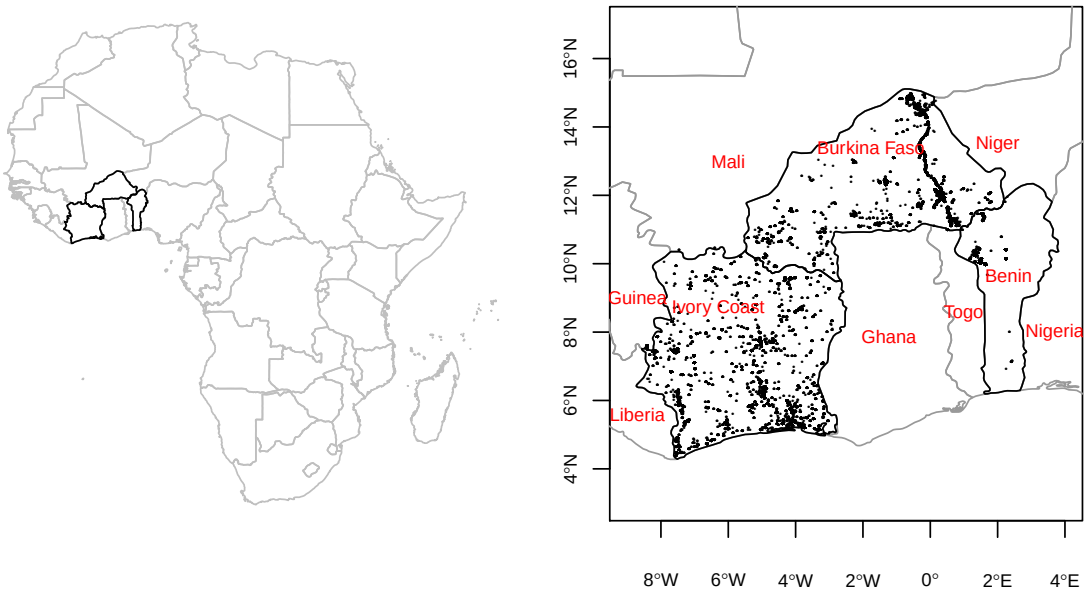

Figure 2.1 – *Left panel: Map of Africa showing in red the countries forming the study area. Right panel: Detailed map of the study area; black dots represent collection localities*

Table 2.1 – *Summary information of the number and density of collection localities, species numbers, total area and area containing information for different grouping factors. Analysis were based on a grid of 10 km^2 size*

Grouping Factors	Area (km^2)	Area with information (%)	N. Collections	Richness Observed	Density mean	Density max
Study Area	730600	13.8	2931	4587	0.40	159
Ivory Coast	330300	18.7	876	3931	0.27	7
Burkina Faso	278800	12.7	1731	1610	0.62	159
Benin	121500	3.2	324	699	0.27	103
Eastern Guinean forest	107600	23.4	373	2958	0.34	12
Guinean montane forest	3000	36.7	17	807	0.57	4
Western Guinean lowland forest	46400	24.6	196	1979	0.43	6
Guinean forest-savanna mosaic	108100	14.1	197	1857	0.18	4
Sahelian Acacia savanna	23700	24.1	549	404	2.32	159
West Sudanian savanna	441800	9.6	1599	2102	0.36	116

of the total area contained information with Benin being the country with least information. Although Burkina Faso has the highest number of collection localities (i.e., 1,731) Ivory Coast is the country with a better coverage of information (Table 2.1). There are six ecoregions in the study area. Although the majority of collections are located in the West Sudanian savanna, this ecoregion contains the smallest coverage (Table 2.1). In contrast, The Guinean montane forest, which is the smallest in area, is better represented, although with the smallest number of collections.

The mean number of observations per collection locality is 18.2; 2,214 collection localities have 18 or less observations (75.5%). The minimum number of observations in a collection

locality is 1 (494 collection localities; 17%) and 1,061 the maximum. There are a total of 4,587 plant species belonging to 1,443 genera and 219 families. There is a high frequency of species with very few recordings and very few species with a high number of recordings. This pattern is the same for genera and families (Figure 2.2). 17.9% of the species (i.e. 823) have only one record. 51.4% (2,360 species) have less or equal 10 records.

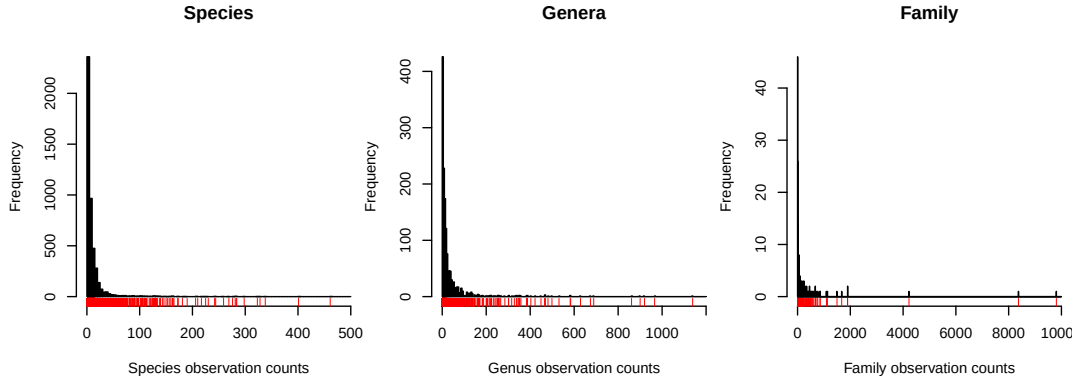

Figure 2.2 – *Frequency distributions of the number of records in each taxonomic level. Red lines (rugs) represent the location of the observations in the histogram*

2.3.3 Bias Factors and Environmental Data

Table 2.2 shows the list of bias factors and environmental layers used to evaluate the causes of bias in the data and the environmental conditions that are over and under-represented respectively. All original layers were prepared and processed using the geographical information system GRASS Version 6.3 (GRASS Development Team 2008). All layers were transformed to UTM coordinates (zone 30N, datum WGS84), scaled to 10 km^2 to match the minimal spatial accuracy of the species collections database and clipped to match the study area.

The climatic data were extracted from the WORLDCLIM database. These data are described in (Hijmans et al. 2005). The data were generated through interpolation of average monthly climatic data from weather stations around the world. The elevation layer was also extracted from the WORLDCLIM database and was included into the geographical information system SAGA (SAGA Development Team 2008) to derive the wetness index variable. The elevation variance was created by calculating the variance of the elevation values using a 9x9 moving window.

Table 2.2 – *List of bias factors and environmental layers used to evaluate the possible sources of spatial bias and the environmental representativeness in the distribution of collection localities. Distance to the coast has a different meaning for Burkina Faso. It represents possible bias in a north-south gradient within the country.*

Layer Name	Derived Layer Name	Abbreviation	Source
Bias Predictors			
Ecoregions of the World	Ecoregions of the World		(Olson et al. 2001)
Main Cities	Distance to Cities		(Defense Mapping Agency (DMA) 1992)
Countries of the world	Distance to the Coast		(Defense Mapping Agency (DMA) 1992)
Rivers	Distance to Rivers		(Defense Mapping Agency (DMA) 1992)
Roads	Distance to Streets		(Defense Mapping Agency (DMA) 1992)
World database on Protected Areas	Protected Areas		(World Conservation Union and UNEP-World Conservation Monitoring Centre 2007)
Environmental layers			
Annual Mean Temperature	Annual Mean Temperature	temp	(Hijmans et al. 2005)
Annual Precipitation	Annual precipitation	prec	(Hijmans et al. 2005)
Temperature Annual Range	Temperature Annual Range	temp_range	(Hijmans et al. 2005)
Elevation	Elevation	elev	(Hijmans et al. 2005)
Elevation	Elevation Variance of elevation	elev_var	(Hijmans et al. 2005)
Elevation	Wetness Index	weti	(Hijmans et al. 2005)

2.3.4 Statistical Analysis

The vascular plant database results from the integration of several sources, specifically of working groups in Burkina Faso and Ivory Coast. Since the aims and purpose of data collection were different from one country to another all statistical analyses were carried out independently for each country and then integrated for the whole area. All analysis were carried out using the statistical software R (R Development Core Team 2009).

Density estimate and Departure from Complete Spatial Randomness (CSR)

Density estimates and departure from randomness of collection localities were investigated from the theoretical backgrounds of point pattern analysis (see Chapter 8 in Cressie 1993). In this study collection localities were considered as the "points" used in point pattern analysis. The first step was to calculate the density or intensity as the number of collection localities per 10 km^2.

To visualize density patterns and as a parameter to include in the final index, a density map of the study area was created using an isotropic Gaussian kernel (Diggle 2003, Baddeley and Turner 2005). An obligated parameter for density estimation is the bandwidth or the smoothing parameter of the Gaussian kernel. The bandwidth was estimated using the method of Berman and Diggle (1989) that looks for the smallest Mean Square Error (MSE) of a kernel estimator (see figure 2.3). 30 km was chosen as the bandwidth value although other values seem plausible given the flatness of the curve.

To quantitatively test if the distribution pattern of the collection localities departed from a complete spatial random distribution (CSR, henceforth), the K-function was used (Schabenberger and Gotway 2005, pp 99-103). Since there are many different environmental

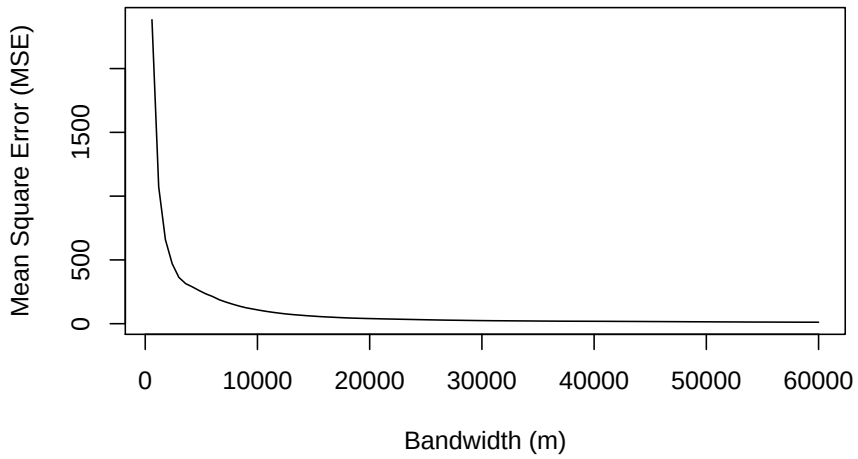

Figure 2.3 – *Relation between the mean square error of the kernel estimator and the bandwidth used to estimate the density of collection localities at different resolutions. Although MSE constantly decrease as the bandwidth increases, the difference becomes negligible at distances greater tan 30 km*

conditions and different processes occurring in the study area, and given that the database is the result of different decisions and sampling procedures varying over time and space, the distribution of the collection localities was considered here as an inhomogeneous point process. To calculate the CSR then, the inhomogeneous K-function was employed following the procedures implemented in Baddeley et al. (2000). Point-wise envelops under CSR were computed by repeatedly making 100 simulations of random distributed points over the study area. Later on, it was checked if the observed pattern (i.e. the one defined by the collection localities) laid inside this envelope.

Bias Analysis

The purpose of the bias analysis was three-fold: (1) to understand what are the factors causing spatial bias in the distribution of collection localities (2) to check if spatial bias of collection localities represent environmental bias as well and (3) to generate a layer representing environmental bias in the study area. The procedures carried out for each of the above points are described below:

Procedure 1: to measure the magnitude of bias in collection localities given the bias factors described in section 2.3.3, each bias factor was divided in four zones based on the range of measured distances in each of them in the study area. Thus, zone 1 represented the area where distances to each bias factor were the smallest while zone 4 areas where distances were the highest. To calculate the size of each zone the fisher algorithm was used (Fisher 1958). This method selects class breaks to group similar values, and at the same time maximizes the difference between classes (Slocum et al. 2005).

Next, bias was quantified for each zone following the index of Kadmon et al. (2004):

$$Bias_d = \frac{n_d - p_d N}{\sqrt{p_d(1-p_d)N}} \tag{2.1}$$

where n_d is the number of collection localities within a specified zone (d), N is the total number of collection localities in the database and p_d is the probability for a given collection locality to be within a zone (d). Since the above equation derives from the normal approximation to the binomial distribution, values become statistically significant when they are greater or less than 1.64 and -1.64 respectively (at α=0.05). Bias values greater than 1.64 represent over-sampled areas, that is, areas with more collection localities than expected from a random sampling design. In contrast, bias values less than -1.64 depicted under-sampled areas.

To estimate p for each zone, the same amount of points as collection localities was generated based on a random sampling design with replacement. The fraction of random points within each zone was taken to be p. The definition of random points and the estimation of the bias index was repeated 100 times. Basic statistics and confidence intervals were calculated.

Procedure 2: If collection localities are biased towards some of the bias factors considered here, modeling species distribution will still not be affected if the geographical arrangement of those bias factors properly represents the environmental variability of the study area. To verify this statement, several steps were carried out. First, the bias factors that showed over-representation of collection localities in any of their four zones were selected. Second, The number of collection localities present in the selected bias factors in the specified zone was counted and the same number of points were created randomly throughout the study area. Third, both sets of points were overlaid with the environmental layers described in section 2.3.3 in order to obtain the values of the environmental variables for each point. Fourth, the frequency distribution of those values were compared using the Kolmogorov-Smirnov test (KS). The KS tests the null hypothesis that the frequency distribution of two samples were drawn from the same continuous distribution (Marsaglia et al. 2003).

Procedure 3: A new layer representing the environmental bias in the study area was created following the same steps as in procedure 1 but using the environmental layers instead of the bias factors. Once the bias index was calculated for each environmental layer and for each zone all layers were summed up to come up with the environmental bias index map. This layer was used as input for the Gap Selection Index.

Database Completeness

To analyze the floristic completeness of the database used in this study, the completeness index proposed by Soberón et al. (2000) was used. This index is based on the comparison

of the total (i.e., estimated) number of species present in a certain geographical area with the number of species observed in the same area: $C = \frac{O}{T}$ where C is the completeness index, O is the observed number of species and T is the total number of species.

The calculation of the C index has to be constrained to a certain geographical area or subdivisions of it, called grouping factor herein. In this study the C index was calculated for the whole study area, for each country, for the WWF ecoregions, and finally for grid cells of different size to identify how the completeness of the database varies with scale.

The observed number of species in each grouping factor was the number of species counted. To estimate the total number of species two non-parametric techniques were implemented:

1. Jackknife first order as a bias reduction method: $S = S_{obs} + L\frac{n-1}{n}$, where n is the number of samples and L the number of species that occur in only one sample (Burnham and Overton 1979, Heltshe and Forrester 1983).
2. Bootstrap: $S = S_{obs} + \sum(1 - p_i)^N$, where p_i is the frequency of species i and N is the total number of collections in the grouping factor (Smith and Belle 1984).

Database Quality Evaluation

We developed the gap selection index as a measure of database quality. For that we considered three factors: the density of collection localities as calculated using the Gaussian smooth kernel (d), the values representing the environmental bias in each country (b) and the database completeness (C). All factors were converted to values between 0 and 1 following the equation of Legendre and Legendre (1998):

$$y_i = \frac{y_i - y_{min}}{y_{max} - y_{min}} \tag{2.2}$$

Then, all factors were subtracted from 1 to ensure that values close to 1 represent deficiencies in data quality. The gap selection index was calculated as:

$$DQI = \frac{d + b + C}{F} \tag{2.3}$$

where F represent the number of factors included in the index. Results of the index are values between 0 and 1, where values close to zero represent areas that have been properly represented, while values close to one represent areas where the density of collection is very low or zero, the information is incomplete and where the environmental conditions are not well represented in the distribution o collection localities.

2.4 Results

Density and Complete Spatial Randomness

The mean density of collection localities in the study area is 0.4 collections per 10 km^2. The highest mean density was found in Burkina Faso (i.e., 0.62) with a maximum of 159 collection localities in a 10 km^2 grid cell. Although the mean density of collection localities in Ivory Coast and Benin is the same, collection localities are better distributed in the first (maximum of 7 collection localities per 10 km^2) while in the second they are distributed in an unique aggregate (maximum of 103 collection localities per 10 km^2). The ecoregion better represented in the database is the Sahelian Acacia Savanna, with a mean density of 2.32 collections per 10 km^2 (Table 2.1)

The spatial distribution of the density estimates for the study area, based on the smoothing gaussian kernel can be seen in figure 2.4. The collection localities are distributed unevenly, with large amount of collections in Burkina Faso and Benin and notably less in Ivory Coast. In addition, collection localities are distributed forming high density patches.

Based on the analysis of the inhomogeneous K-function, the distribution of collection localities is not random. The same statement applies for each country (Figure 2.5). Contrary, the pattern in each country follows a clustered distribution, which is less accentuated for Ivory Coast. The observed spatial clustering or aggregation of collection localities is present even after allowing for spatial variation in density.

Bias Analysis

In general all bias factors have a strong influence on the spatial distribution of collection localities in the study area. For Ivory Coast, there is a clear over-representation of collection localities in zone 1, that is, in areas close to each of the bias factors but most importantly to the vicinity to cities, to the coast and to streets. In zone 2, there is almost no bias but in zones 3 and 4 the trend is towards an under-representation of collection localities (Figure 2.6). Closeness to streets and specially to protected areas are the factors explaining the over-representation of collection localities in Burkina Faso. Also in Burkina Faso there seems to be a preference to collect far away from the main cities, the coast and streets. At intermediate distances from all factors a negative bias seems apparent (Figure 2.6). As for Benin, in places situated close to rivers and streets, an over-representation of collection localities is found. At intermediate distances there is also a collection over-representation regarding protected areas, cities and the coast and at long distances for all bias factors a negative bias exist (Figure 2.6).

In general, a positive bias or an over-representation of collection localities in some regions of the study area also represent an environmental bias. That means, some environmental

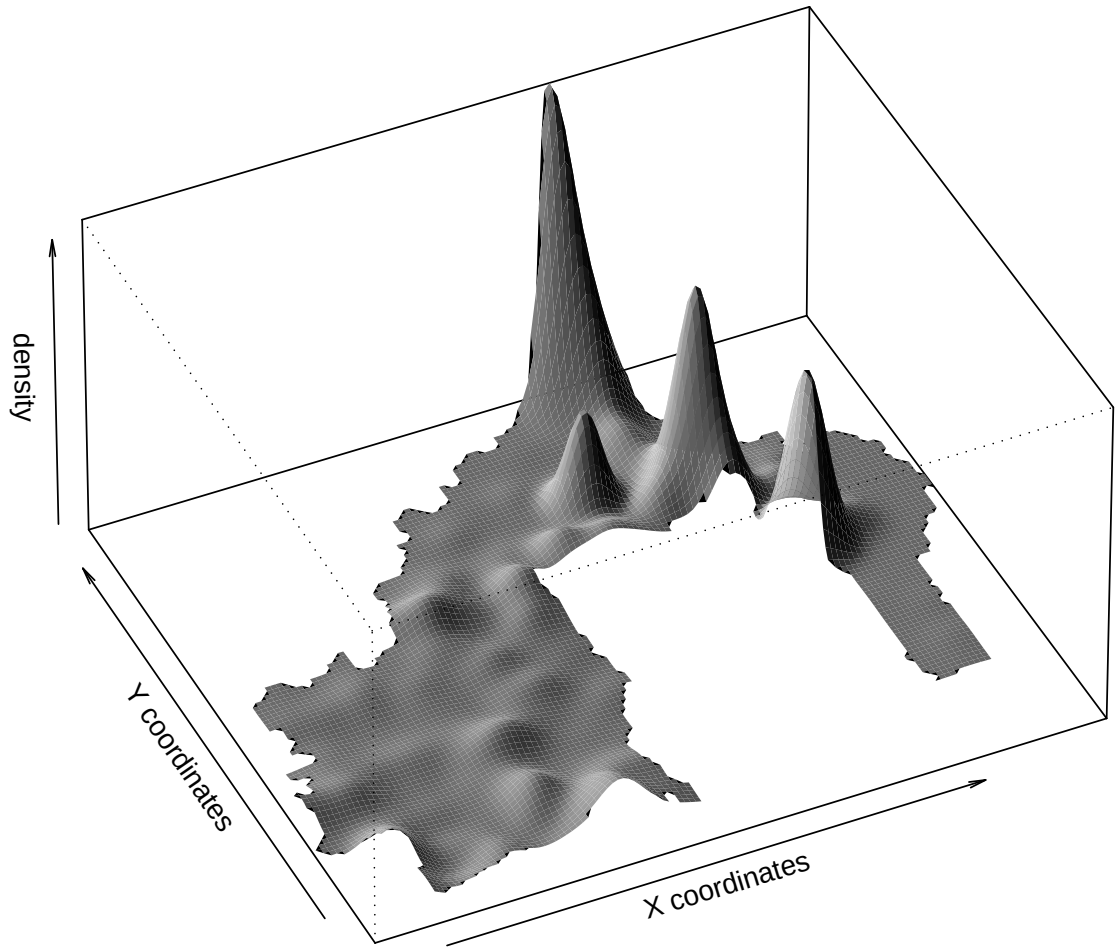

Figure 2.4 – *Three dimensional view of collection localities density patterns estimated based on a smoothing Gaussian kernel*

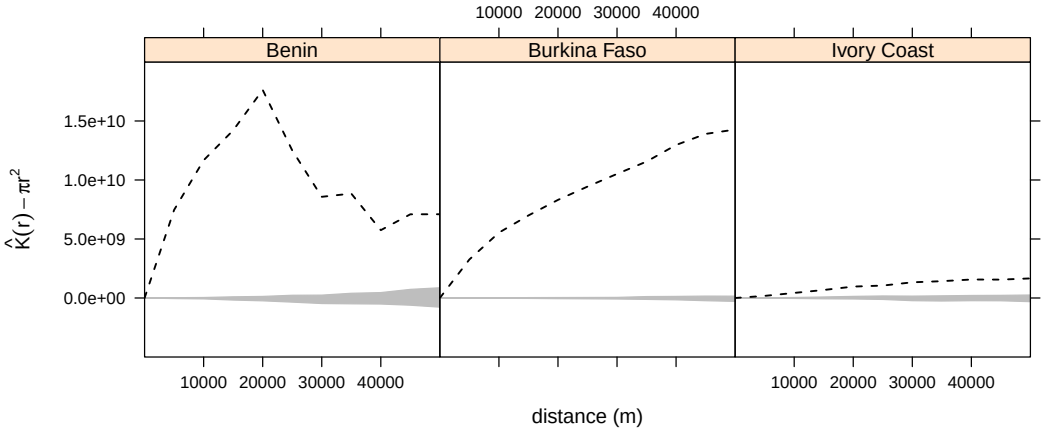

Figure 2.5 – *Point pattern estimates of the collection localities based on the inhomogeneous Ripley's K-function. Displayed are envelopes (gray) representing the area occupied by realizations of 100 simulated random patterns. Black dash lines are the estimated K values of the collection localities for different distances. The line is expected to be inside the envelope if the pattern of collection localities is random. Above the envelope a clustered patter is represented.*

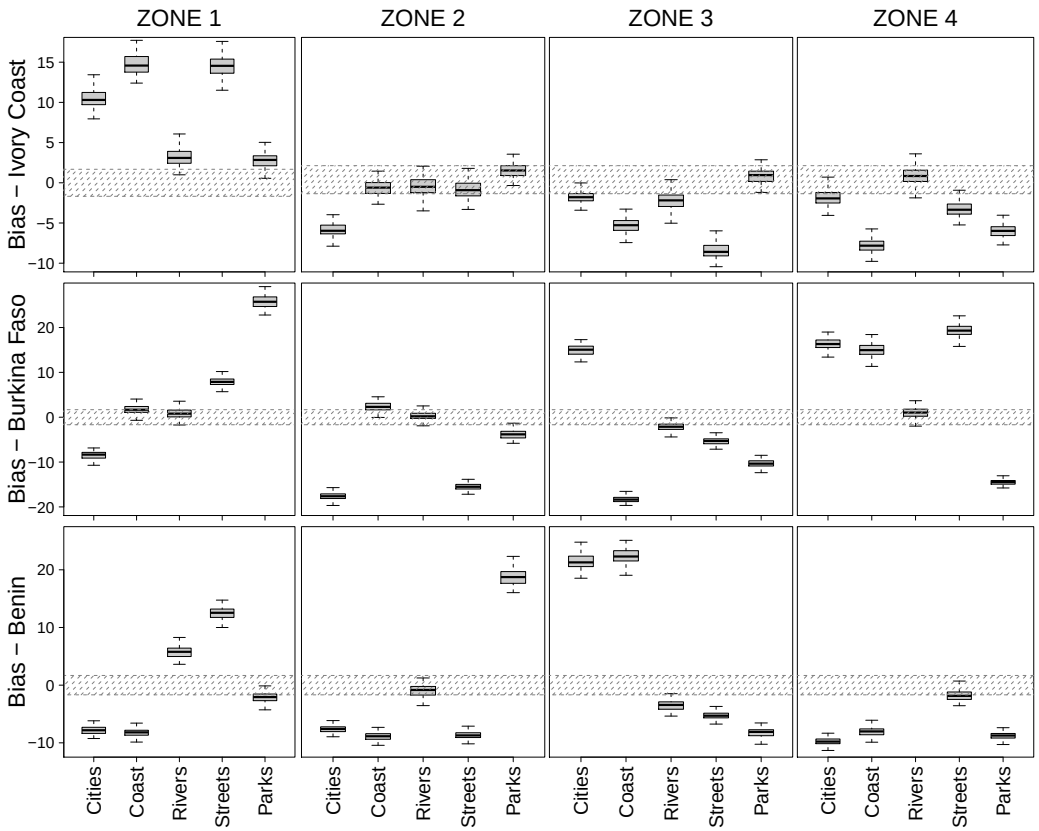

Figure 2.6 – *Bias estimates (as calculated from equation 2.1) for each country (rows) in each distance zone (columns) and for each of the bias factors considered in this study (i.e., distance to cities, to the coast, to rivers, to streets, and to natural parks). Zone 1 represents the range of distances closer to each of the bias factors and zone 4 the range of distances farthest apart from them. Shadow polygons represent the range of values where no bias is expected. If boxplots are within this area than the number of collection localities are as expected from a random sampling scheme (i.e., no bias). Boxplots above and below this area represent over or under-sampling respectively.*

conditions are over-represented while others are under-represented in the distribution of collection localities. If the distribution of collection localities were not environmentally bias, it would be expected to find non-significant differences between the frequencies of environmental values for those in the location of collections in the database and those in randomly selected points. Non-significant differences were only found in Ivory Coast, where the fact that the majority of collection localities are near cities and rivers does not imply an environmental bias considering the variance in elevation (elev_var) and the wetness index (weti); in all other cases the differences are significant (Table 2.3).

Visualization of differences between environmental value frequencies for the locations of collections previously identified in over-sampling zones and random distributed locations in the study area, allows a better understanding of the environmental conditions that have been under or over-represented (i. e. bias). An example of such comparison can be seen

Table 2.3 – *Results of the Kolmogorov-Smirnov test by comparing the environmental values of collection localities within the zones and bias factors where they were over-represented and the environmental values of the same number of points located randomly over the study area. Environmental bias is found in those areas were differences are significant. Significance coding: * $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$;* ns *non significant*

	Environmental variables	temp	prec	elev	elev_var	temp_range	weti
Ivory Coast							
ZONE 1	distance cities	0.34 ***	0.18 ***	0.44 ***	**0.05 ns**	0.48 ***	**0.06 ns**
	distance coast	0.11 *	0.45 ***	0.7 ***	0.13 **	0.68 ***	0.18 ***
	distance rivers	0.14 **	0.24 ***	0.34 ***	**0.06 ns**	0.27 ***	0.36 ***
Burkina Faso							
ZONE 1	distance streets	0.24 ***	0.39 ***	0.22 ***	0.19 ***	0.32 ***	0.11 ***
	distance protected areas	0.5 ***	0.49 ***	0.4 ***	0.21 ***	0.51 ***	0.18 ***
ZONE 3	distance cities	0.29 ***	0.26 ***	0.23 ***	0.16 ***	0.18 ***	0.19 ***
ZONE 4	distance cities	0.81 ***	0.81 ***	0.45 ***	0.34 ***	0.77 ***	0.27 ***
	distance coast	0.91 ***	0.89 ***	0.36 ***	0.42 ***	0.93 ***	0.34 ***
	distance streets	0.87 ***	0.89 ***	0.45 ***	0.49 ***	0.85 ***	0.3 ***
Benin							
ZONE 1	distance streets	0.52 ***	0.73 ***	0.83 ***	0.74 ***	0.36 ***	0.42 ***
ZONE 2	distance protected areas	0.52 ***	0.74 ***	0.8 ***	0.73 ***	0.44 ***	0.57 ***
ZONE 3	distance cities	0.45 ***	0.64 ***	0.71 ***	0.67 ***	0.35 ***	0.39 ***
	distance coast	0.57 ***	0.7 ***	0.82 ***	0.67 ***	0.37 ***	0.4 ***

in figure 2.7

The map in figure 2.8 depicts the sum of the bias estimates for each of the environmental variables used in this study (see Table 2.2). Clearly, environmental conditions in areas near the coast in Ivory Coast, in and around the eastern Guinean forest in Benin and the Sahelian zone in Burkina Faso have been over-represented. In contrast, wide extensions of the savannas and the forest-savanna mosaic have been under-represented.

Completeness Analysis

A general comparison between the two non-parametric techniques employed indicated that results of the Jackknife 1 estimator are in general higher than results of the Bootstrap estimator and therefore, completeness values were always higher when calculated based on the Bootstrap technique.

Estimates of species richness and completeness were calculated for different grouping factors (Table 2.4). In general, the floristic knowledge of the study area is good, as shown by the high values of the completeness index. Comparing the three countries independently, Burkina Faso is the less studied country since it has the lowest completeness value. From 1610 plant species observed so far, there can be between 531 to 236 species still not described in the database. Regarding the WWF ecoregions, all except for the Guinean montane forest have been properly studied. From 807 species registered in the database, a maximum of 1419 or 1057 species have been estimated and according to the Jackknife 1 and bootstrap species richness estimators, which results in completeness values ranging

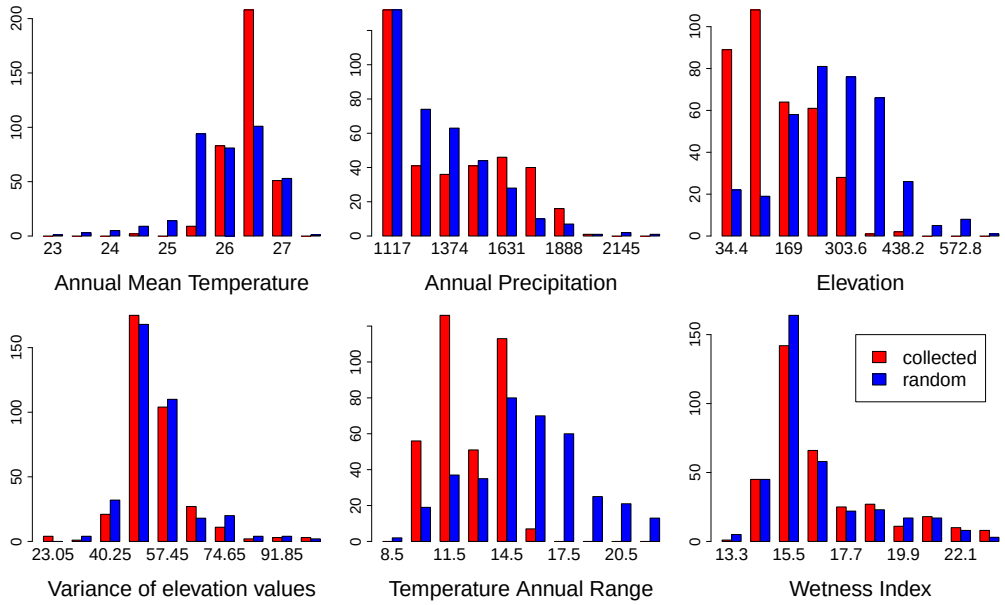

Figure 2.7 – *Example of the difference between the frequency distribution of environmental values found for all collections located in areas close to cities (i.e., ZONE 1) in Ivory Coast and for the locations of randomly distributed points in the study area. No significant differences exist for the variance of elevation (elev_var) and wetness index (weti). On the contrary, for all other environmental variables the differences are significant (see also table 2.3). There is an over-representation of low elevation areas while areas of high altitude have been under-represented. The same case applies for temperature annual range (temp_range) and the opposite for annual mean temperature (temp) and annual precipitation (prec).*

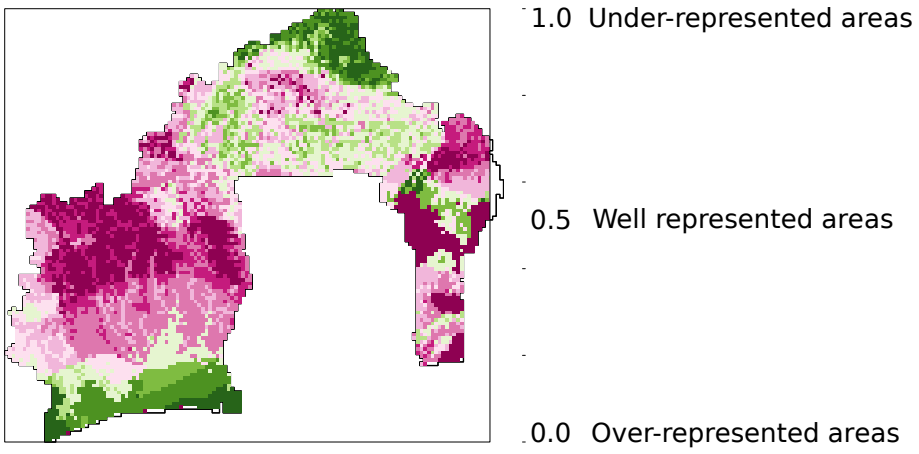

Figure 2.8 – *Representation of environmental bias in the study area. Values close to one (green) represent areas where environmental conditions are under-represented. Areas assigned values close to zero (white) have been visited as expected by applying a random sampling scheme. Environmentally over-represented areas in the distribution of collections localities are those with values close to one (purple).*

from 0.57 to 0.76 respectively. In contrast, the ecoregion so far best represented in the database is the Eastern Guinean Forest, having the highest completeness values.

Table 2.4 – *Analysis of species richness and completeness estimates. Richness observed are the total number of species counted in each of the grouping factors. Richness estimates were calculated by using two non-parametric estimation techniques (i.e., Jackknife 1 and Bootstrap). Completeness is calculated by dividing richness observed by richness estimates.*

Grouping Factors	Richness observed	Richness estimates		Completeness	
		Jackknife 1	Bootstrap	Jackknife 1	Bootstrap
Study Area	4587	5409.7	4989.2	0.85	0.92
Ivory Coast	3931	4601.2	4273.4	0.85	0.92
Burkina Faso	1610	2141.7	1846.3	0.75	0.87
Benin	699	853.5	775.2	0.82	0.90
Eastern Guinean forest	2958	3703.9	3329.0	0.80	0.89
Guinean montane forest	807	1419.7	1057.1	0.57	0.76
Western Guinean lowland forest	1979	2800.7	2354.5	0.71	0.84
Guinean forest-savanna mosaic	1857	2617.1	2207.9	0.71	0.84
Sahelian Acacia savanna	404	534.7	464.6	0.76	0.87
West Sudanian savanna	2102	2796.5	2409.6	0.75	0.87

Completeness analysis were also applied on a grid cell basis. Different cell sizes (i.e., resolutions) were used (i.e., $10\,km^2$, $30\,km^2$, $60\,km^2$, $120\,km^2$). As expected, correlations between number of species observed and estimated were very high at all resolutions (in all cases a correlation coefficient equal 0.99). On the contrary, correlations between estimated species richness and completeness values were low (Figure 2.9). In general, grid cells with the highest number of species are not necessarily complete. Complete cells are in Benin and Burkina Faso although the two countries are less studied in comparison to Ivory Coast (Table 2.4). Note that the majority of the grid cells in Benin have a completeness index close to zero or zero.

The percentage of grid cells containing information increases with an increase in cell size (Table 2.5). That means, at a 10 km^2 resolution the percentage of cells with information (13.8%) is smaller than the number of cells with information (85.1%) at a $120\,km^2$ resolution. At lower resolutions there is more information available for analysis resulting in a better coverage of the study area.

As a result of the clustered distribution pattern of collection localities, there are few areas with high density and most of the remnant area has either non or a very small density of collection localities. Consequently, there are either areas with high completeness index values or areas with very low completeness values. Intermediate completeness values (i.e., between 0.2 and 0.6) do not exist (Figure 2.10). However, the percentage of grid cells with a completeness value equal or higher than 0.6 increases until 60 km^2. There is also a constant increase of grid cells with completeness values higher than 0.8 as the resolution increases (Figure 2.10).

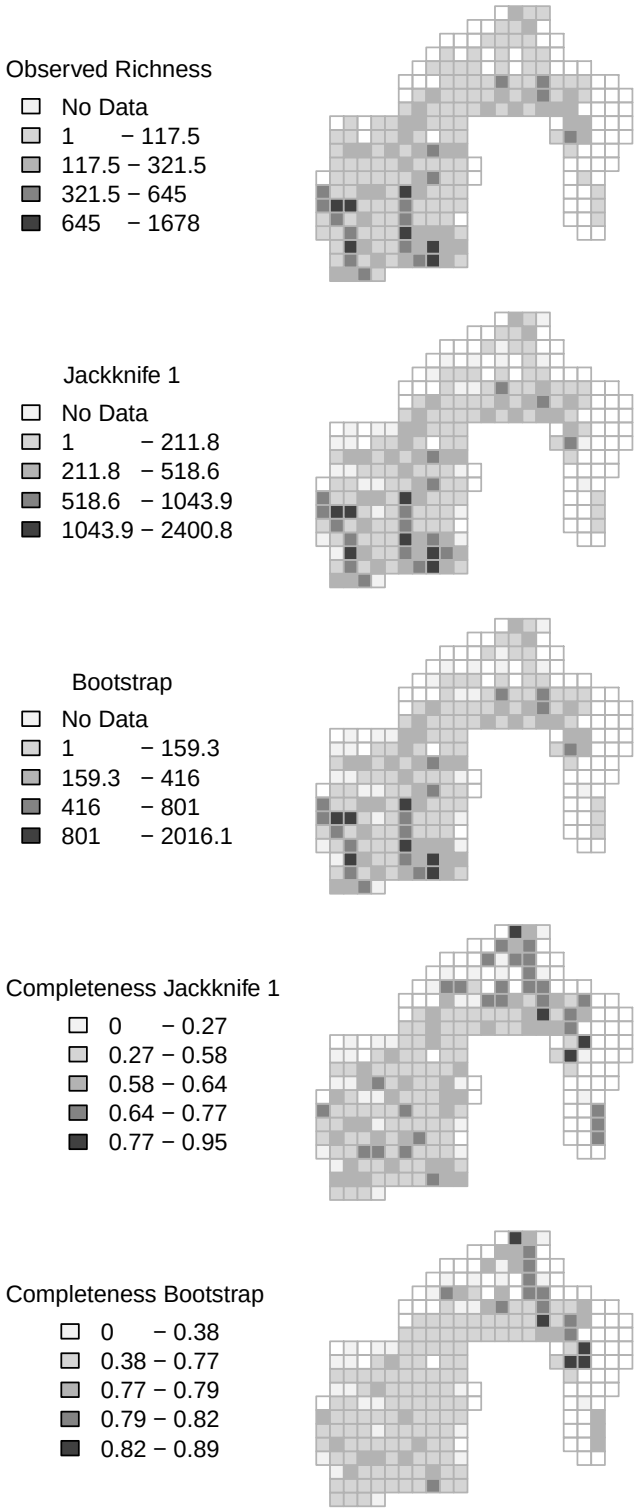

Figure 2.9 – *Maps of observed species richness, estimated species richness as calculated using two non-parametric estimation techniques (i.e., Jackknife 1 and Bootstrap), and Completeness Index (i.e., richness observed divided by richness estimates). All illustrations are based on the analysis done at a 60 km^2 resolution. The Jackknife 1 estimator produced in all cases higher species richness estimates while the Bootstrap produced more conservative numbers.*

Table 2.5 – *Differences between the number and percentage of grid cells containing no information at different spatial resolutions.*

Resolution (km^2)	N. Cells	Cells with no information	%
10	7306	6295	(86.2)
30	884	444	(50.2)
60	247	65	(26.3)
120	74	11	(14.9)

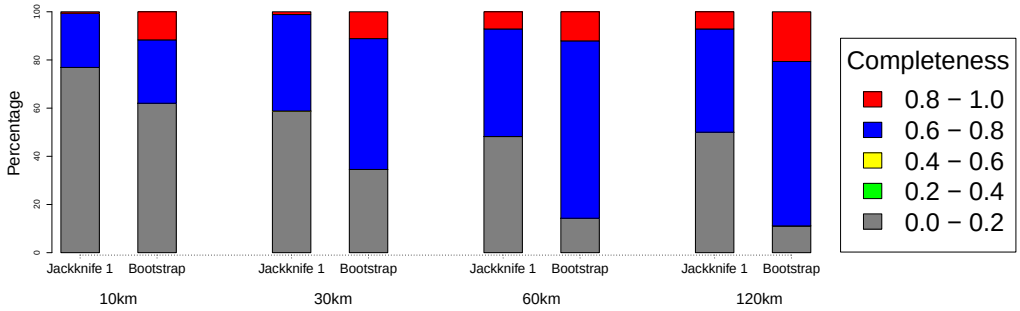

Figure 2.10 – *Barplots showing the percentage of grid cells in five completeness classes (see legend). Calculations were done for four different spatial resolutions and considering the two non-parametric techniques used for species richness estimations (i.e., Jackknife 1 and Bootstrap) at each resolution.*

Gap Selection Index

Based on the density distribution of collection localities (Figure 2.4), the degree of environmental bias (Figure 2.8) and the floristic completeness of the database (Figure 2.9), the gap selection index was calculated (Figure 2.11). In the index, values close to zero represent areas that have been well studied, actually, where the density of collection localities is high, where the environmental conditions have been properly represented and where the floristic information is complete. Four spots in the study area fulfill these conditions: the Sahelian zone and surroundings of Comin-Yanga city in Burkina Faso, areas close to the coast and in the border with Liberia in Ivory Coast and the region of the eastern Guinean forest in Benin (Figure 2.11).

From a pessimistic point of view, problematic zones can be considered as those having values greater than 0.8. 71.1% of the total area are within this zone. 70.9%, 64.9% and 86.2% of the area in Ivory Coast, Burkina Faso and Benin respectively have values greater or equal 0.8. The ecoregions better represented in the database are the Sahelian Acacia savanna and the Guinean Montane forest with only 8% and 13.8% of their areas in need of more information. On the contrary, the West Sudanian savanna and the Guinean forest-savanna mosaic are almost without information with 78.9% and 92.2% of their area respectively with index values greater than 0.8.

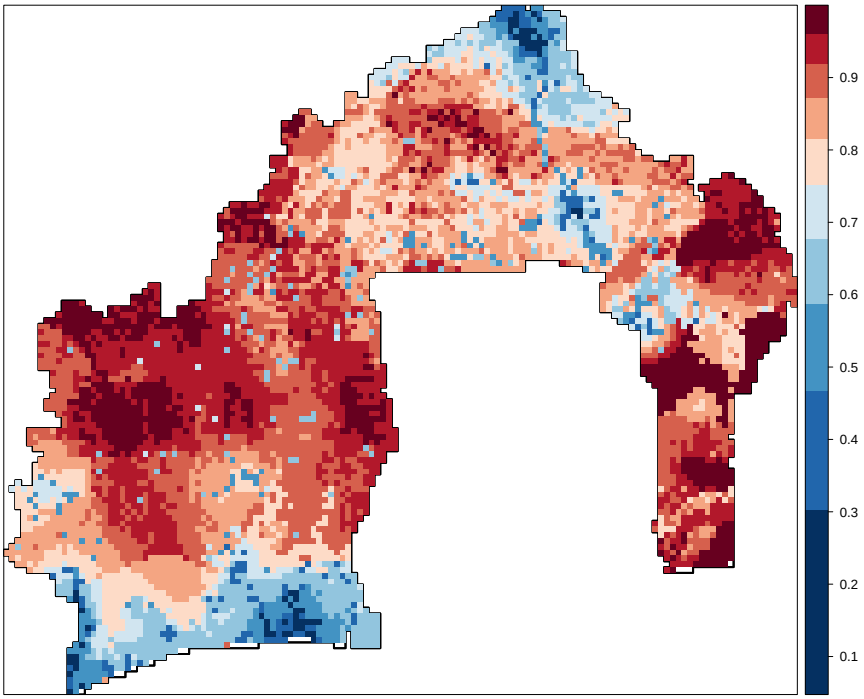

Figure 2.11 – *Map of the Gap Selection Index (GSI). The main goal of the index is to emphasize those areas that have been poorly visited and contained environmental information not well represented by the distribution of collection localities in the database. Therefore, values close to one represent under-represented areas while places with values close to zero have received enough attention and have been well studied. The index has been calculated by integrating information on collection densities, environmental representativity and floristic completeness on a pixel based approach (i.e.,10 km^2)*

2.5 Discussion

Uneven efforts of plant collection in West Africa

The database used in this study is the result of more than 6 years of collection efforts. A great part of the data come from field activities, other from geo-referenced data in herbarium collections or regional atlases. There are a total of 4587 species in the database and the completeness estimates in table 2.4 suggest that there is a good floristic knowledge of the region. However, the average density of collection localities per 10 km^2 is very low although in some specific places the density reaches 159 collection localities per 10 km^2. That is in agreement with the marked cluster pattern of collection localities observed.

Although the database has been already used for estimating patterns of plant diversity in the region (Schmidt et al. 2005, Thiombiano et al. 2006), this has not necessarily been the main goal motivating the construction of the database. Instead, many of the data have been generated from specific projects with their focus on specific areas. In Burkina Faso for example, a special focus has been given to the Sahelian acacia savanna ecoregion

where a high density of collection localities is found. Specific macroecological analysis of this particular region have taken place (Schmidt et al. 2008). Other "hot-spots" of plant collections in Burkina Faso are situated in small areas that have been the focus of investigation by students and researches in the region. The same situation occurs in Benin, where only a special area has been the focus of research, that is, the Eastern Guinean Forest (see Figure 2.4), where the maximal density found is 103 collections per $10\,km^2$.

Not surprisingly, the south of Ivory Coast, according to the results of this study, is the place that has been better studied, since in this area is where the well known Guinean Forest diversity hotspot is located (Myers et al. 2000), one of the areas with high species richness and endemism. Several studies have been carried out to investigate different aspects of the composition, structure and dynamics of the forest ecosystems in this area (Chatelain et al. 2004, Nussbaumer et al. 2005). However, in comparison with Burkina Faso and Benin, the coverage of collection localities in Ivory Coast is better distributed across the country although still showing a clustered pattern.

2.5.1 Bias: the reiterative issue

Spatial bias in biological databases is one of the most repeatedly mentioned issues in biogeographical research.In addition, it is presumed as one of the factors potentially harming the results of biogeographical analysis (Funk and Richardson 2002), but its evaluation as well as its influence on model output is normally not explicitly made. However, it has been demonstrated that spatial bias can have a big influence on model outcome and performance as well as in the establishment of the effect of environmental variables on the defined niche of a species. For that reason, we agreed with other authors that a very first step, before modeling species distribution, is to explicitly evaluate the database in terms of spatial bias and to understand the possible causes that led to that bias.

We used a similar approach to both identify the factors influencing spatial bias in the database and estimate environmental bias as implemented earlier by Kadmon et al. (2004) and Loiselle et al. (2008). Although Kadmon et al. (2004) found significant differences between the distribution of collection localities and that of the rainfall conditions based on a random selection of localities in the study area, they demonstrated that predictions of habitat suitability were not biased, since the statistical difference was weak (although significant). In this study, strong statistical differences were found (see 2.3) and therefore it is concluded that model predictions based on the database as it is and at high resolution (i.e., $10\,km^2$), will produce biased and false estimates of habitat suitability or species ranges predictions.

What is the appropriate scale of analysis?

Finding the appropriate scale of analysis is one of the most controversial and studied issues in ecology (Hurlbert and Jetz 2007). The scale or resolution of analysis should comply with the inherent properties of any given data set (Hengl 2006). For the first time a dataset consisting of the distribution of almost all plants in West Africa with a spatial accuracy of 10 km^2 has been compiled. At the same time, the availability of environmental information derived from remote sensing and geographical information systems provides a good opportunity to carry out analysis at local scales and with a high spatial resolution. The question here is if 10 km^2 resolution is the proper scale for macroecological analysis?

Several of the analysis carried out in this study indicate that 10 km^2 is not necessarily the best scale of analysis. First of all, the mean square error of the bandwidth calculated to estimate the density (Figure 2.3) is higher at 10 km and diminishes at longer distances with small differences above 30 km. That means that if we calculate the density patterns with a bandwidth of 10 km, the final estimates will have a bigger error than at longer distances. Hengl (2006) have recommended the inspection of the density of a point pattern as one of the criterion to define the right pixel size.

Secondly, it is clear from table 2.5 and figure 2.10 that the percentage of grid cells containing information increases as the resolution increases and that the amount of grid cells with complete information also increases as the cell size increases. For further analysis, for example estimating species richness patterns, having more information that covers more of the study area would generate better and more reliable results. For this study case and based on the results presented here, a pixel size of 60 km^2 is recommended as the proper scale for analysis.

Disentangle the Site Selection Index

It is not the first time that a methodology is developed to identify areas where information is missing. For example Funk et al. (2005) developed a methodology, which they called survey-gap analysis, to identify the location of future collection activities. For that purpose, they used in conjunction a set of environmental variables to derive an environmental diversity (ED) measure (see Faith and Walker 1996) and the set of collected sites, which they integrated into a complementary analysis to select sites that would contribute new taxa.

The novelty of the gap selection index concept developed in this study is the integration of different independent criteria, which makes the selection of target sites objective and efficient. Relying on just one criterion makes the identification of target sites unpractical. For example, if some particular places are to be visited based on density estimates, then

areas with low density of collection localities will be chosen. Obviously, information is still missing in those areas. But if the environmental conditions characterizing those areas are very similar to areas where collection density is high, then similar vegetation structure and floristic composition will be expected and no new data will be added to the database. A more efficient use of the resources at hand, will be to go to places that combine low collection densities and underrepresented environmental conditions

Another important combination of criteria for site selection refers to collection densities and database completeness, compared on a grid cell basis (Soberón et al. 2007). The expected behavior of the relationship between these two criteria is an increase in completeness with an increase in collection density. In this study, this relationship is weak (Figure 2.12). In general, the majority of grid cells have low density values and yet complete. In conclusion, a good estimation of the floristic composition of the study area requires few collection localities properly distributed.

Areas also interesting to visit are those where the density is high but the completeness is low. It means that potentially there might be a set of new species to find, probably range restricted species whose detectability is difficult. This is not necessarily the case in the database presented here. Grid cells with high density values are all relatively complete (Figure 2.12).

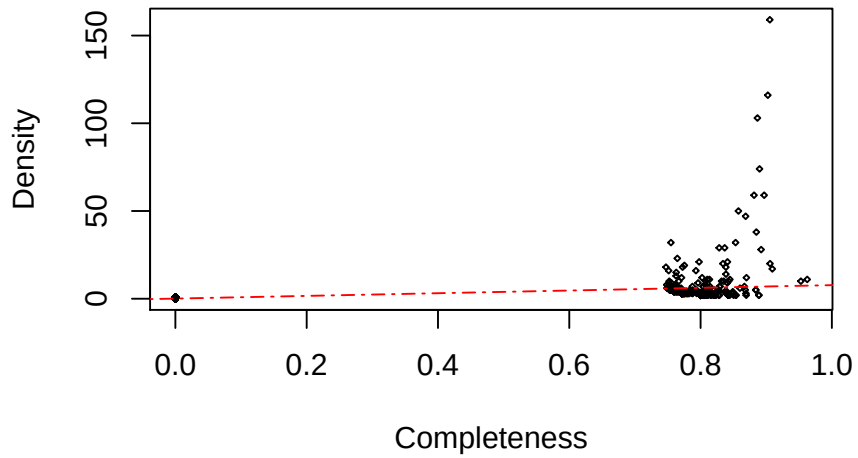

Figure 2.12 – *Relation between density of collection localities and completeness values per 10 km^2. Red line represent the best fit of a simple linear regression model.*

One of the main purposes of the gap selection index is to identify areas where future sampling should be focused on in order to have a more complete picture of the floristic composition of the region.

Still, large areas are unvisited and going to all of them is unrealistic. In addition, many areas might be strongly affected by human influence. The use of additional tools that help in the effective selection of target areas is needed. One of these tools which can served this purpose is Google Earth ®. There are already efforts done in making use

of the integration of GIS analysis and visualization on Google Earth for scientific and communication purposes (e.g. Conroy et al. 2008)

An example of the gap selection index map displayed in Google Earth is shown in figure 2.13. The contrast between forested and deforested areas is very clear in the high resolution image. Still, the whole area has been assigned the highest values of the gap selection index. Without any doubt, a visualization of this type could help in both understanding the processes currently going on and identifying areas that really deserve more investment regarding field work and resources.

In this study we decided to use non-parametric techniques to estimate species richness based on the species observed in the different collection localities. Specifically, we used the first order Jackknife and the Bootstrap estimators. The first one has been constantly ranked within the most precise techniques and the second one is always considered as a technique that constantly underestimates the real value. It was because of those properties that the two techniques were selected, so a range of possible values could be given and compared. Also, because their computational implementation make their calculation feasible. However, there is a range of species richness estimator techniques that have been successfully used for the same purpose as here. For example, Baselga and Novoa (2006) and Jiménez et al. (2009), used rarefaction curves to estimate species richness values and compared them to the observed species richness to evaluate the completeness of their databases.

Modeling Plant diversity in West Africa: where from now?

After analyzing the quality of the database as done so far, the next question is: is it possible to use the information in the database to model plant diversity patterns in west Africa? From the gap selection index (Figure 2.11) is clear that a considerable part of the study area is missing information and is not well represented in the database. Although several techniques can potentially be applied to model diversity patterns, it is important to recognize that there will be a considerable amount of uncertainty present in predictions, especially in those areas missing information. It is recommended that any efforts to display plant diversity patterns should be accompanied by the gap selection index map as a representation of the uncertainties of the outcomes of any modeling approach.

Another aspect that has to be taken into consideration when estimating diversity patterns, regardless of which technique is being used, is the spatial autocorrelation of the collection localities. This is the result of the clustered pattern of the collection localities in the database (Figures 2.4 and 2.5). Both, accuracy of species diversity predictions and estimation of environmental determinants of species diversity will be affected and wrongly calculated if spatial autocorrelation is not considered (Dormann 2007a)(but see Dormann

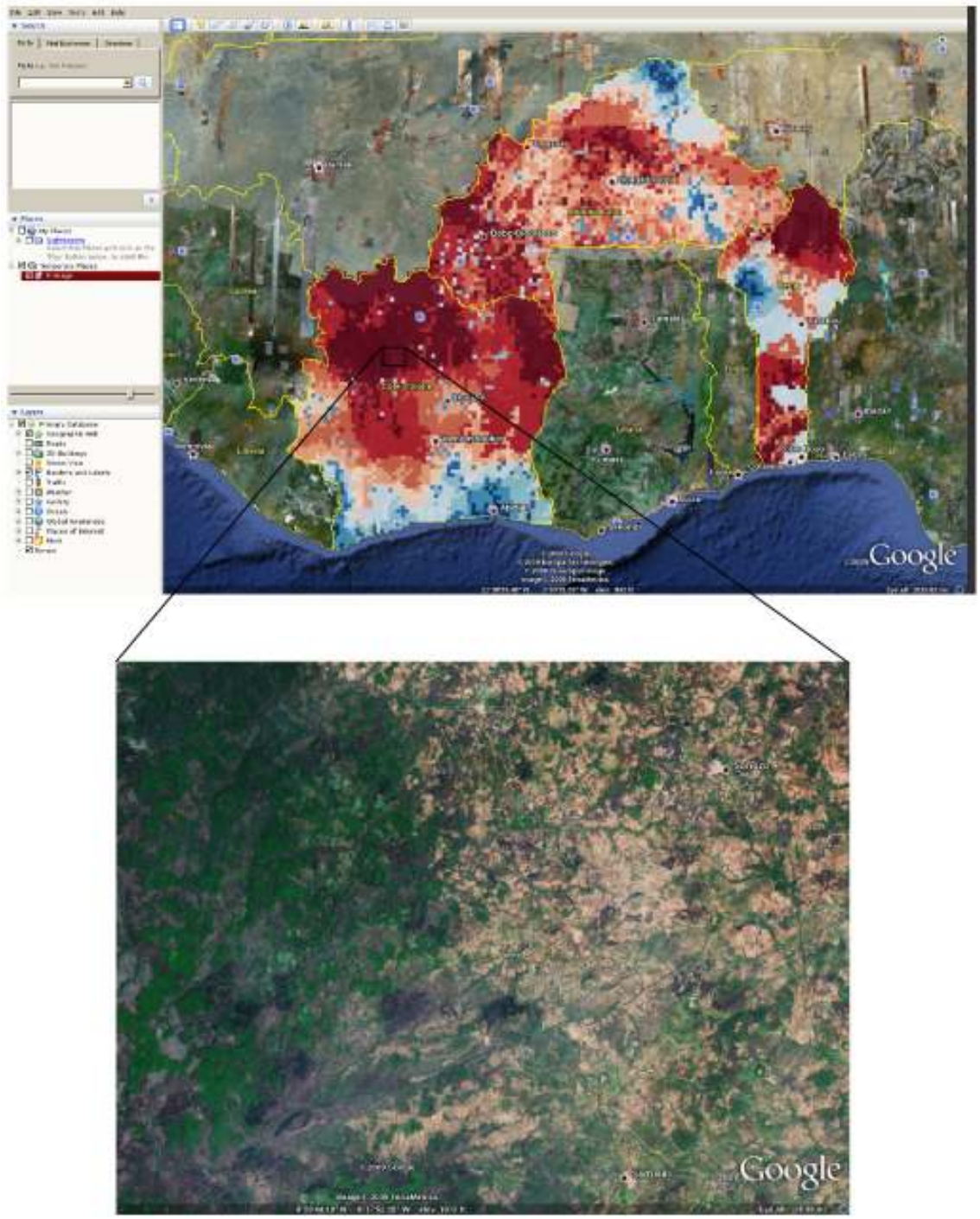

Figure 2.13 – *Example of importing and displaying the gap selection index map to Google Earth. Below is an image with high resolution of a region which has been cataloged of highest priority*

et al. 2007, for a review of different methods to deal with spatial autocorrelation). Based on our database, a generalized linear model was fitted to the species richness counts per grid cell and the environmental variables described in table 2.2. To identify spatial autocorrelation the Moran's Index was calculated to the residuals of the model and the final correlogram can be seen in figure 2.14 where the spatial dependence of species richness counts at smaller distances is evident.

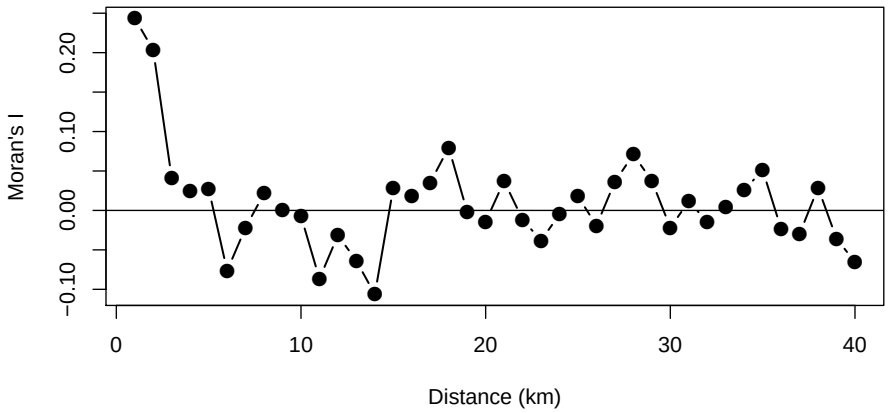

Figure 2.14 – *Spatial correlogram using the Moran's I index of the residuals of fitting a generalized linear model of species richness counts and the set of environmental variables used in this study*

Many techniques can be used to predict and model plant diversity patterns based on our database. All these techniques can be grouped into two main modeling approaches. In the first group are all techniques that directly relate species diversity (e.g. species richness counts) and environmental variables. From the relationships found (i.e. variable coefficients), models are able to predict species diversity to the total extent of the region of interest. Within this group are many of the regression-like model techniques and other approaches like neural network and Bayesian models. Algar et al. (2009) labeled these models empirical diversity theory approaches, since they are constantly used to test macroecological hypothesis that seek to explain the causal factors determining diversity distribution patterns (e.g. Kreft and Jetz 2007).

A common approach to model species diversity, used also in similar studies, is to select only those grid cells that are above a selected completeness threshold for modeling (e.g. Romo et al. 2006). Although information might be lost following this approach, the reliability of predictions increases since the modeling itself is being made with good data. Another possibility is to use all cells with data, so no information is lost, and weight those cells with the completeness index calculated for each of them.

The second group of modelling approaches is the species niche modelling (see Elith et al. 2006, for a review and performance comparison of different methods). The principle of this approach is to model individual species separately and create species range map for each of them. Afterward, all maps are added together to create a final map of species

richness. A major constraint of applying this approach to the database is that a great amount of species have been collected few times (see figure 2.2). It is a demonstrated issue that the accuracy of the models is positively related to the sample size (Stockwell and Peterson 2002). If the recommendation of using species with 10 or more sampled occurrences given by Hernandez et al. (2006) is followed, than only 1423 species will be considered for analysis, that is, only 31% of all species. However, even 3 occurrence localities have proved to be useful to model species ranges (Pearson et al. 2007).

Again, dealing with spatial bias (i.e. spatial autocorrelation) in the distribution of the collection localities becomes an issue when using this approach. Some research has been done to deal with this issue (e.g. Phillips et al. 2009, Kadmon et al. 2004, Allouche et al. 2008, De Marco Jr et al. 2008) and the methodologies recommended can potentially be also applied for the database. Algar et al. (2009) made a comparison of the two main approaches described above to estimate species richness patterns and predictions to the future, and found that after dealing with the spatial autocorrelation issue, the first approach (i. e. empirical diversity theory approaches) did a significant better job.

Conclusions

A lot of work, money, and effort have been invested in the creation of the databases forming the basis for this study. Scientific investigations already done are a good example of utilizing these data for research and conservation applications (Schmidt et al. 2008; 2005, Thiombiano et al. 2006). However, the use of the database for macroecological studies at regional scales might be limited by a series of factors. The distribution of collection localities has not been done in a manner expected in statistical techniques. Particularly, collections do not follow a random distribution but rather a very clustered pattern. There are few areas that have been well investigated, but information is still missing for most of the extent of the study area.

It was demonstrated how the correlation between several bias factors and the distribution of collection localities is very high. To mention few examples, researches have conducted their work near the coast and near to the main streets in Ivory Coast while in Burkina Faso the Sahelian zone has received extra attention. Unfortunately, this collection distribution bias represents environmental bias as well. Many areas with specific environmental conditions have not been visited yet, and their inclusion into models that seek to predict species distribution ranges to those areas will result in false estimates.

If biogeographical applications based on the database are expected in the short term, it is strongly recommended to find the proper modeling techniques that are robust to spatial bias in the distribution of collection localities. Several approaches and modeling techniques have been tested to deal with this issue (see Kadmon et al. 2004, Phillips et al. 2009, for

example) with positive results. Nonetheless, if there are funds and resources to organize field campaigns, targeting the areas identified in the gap selection index will fill up data gaps and will decrease the amount of bias currently present in the database.

CHAPTER 3

DEALING WITH COLLECTION RECORDS BIAS IN SPECIES NICHE MODELING

Without wanting to be elitist, the thing that will prevent literate programming from becoming a mainstream method is that it requires thought and discipline. The mainstream is established by people who want fast results while using roughly the same methods that everyone else seems to be using, and literate programming is never going to have that kind of appeal. This doesn't take away from its usefulness as an approach.

Patrick TJ McPhee

3.1 Abstract

Niche models are tools currently used to investigate a range of questions and hypotheses in ecology, evolution, and biogeograhy. The main input to niche models are geo-referenced locations of the target entity (e.g., species). However, final predictions will be inaccurate if there is spatial bias in the way collection records were taken. In this study we applied a niche model using three different approaches to evaluate both differences in model performance and ways to correct model predictions when the input data is spatially biased We also investigated the resultant species richness patterns derived from each approach. We used the maximum entropy technique (as applied in Maxent) to model the probability of occurrence of 1423 vascular plant species in West Africa (i.e., Burkina Faso, Ivory Coast and Benin). Maxent requires as input species occurrences and background data. We developed three sets of background data, random sets (default) and two sets (target and index background) to counteract the influence of bias in collection sampling. Vascular plant richness was estimated by converting probability estimates to binary responses and adding up individual species predictions. In average, all model prediction approaches perform well (AUC > 0.8). However, target background based models achieved better model performance. Visually, only target based models corrected for bias in collection localities. Richness patterns of vascular plants varied considerably and estimates based on target background models were more accurate to the known positive north-south richness gradient in the study area. Correcting for spatial bias in collection records is a necessary step to improve model performance and prediction accuracy. The use of target background demonstrated to be a suitable option for the type of data presented here, where the spatial arrangement of collection localities is strongly biased.

3.2 Introduction

Currently, species distribution models (SDMs) are being extensively used in different applications in ecological research, ranging from testing ecological theories (Graham et al. 2006), guiding field surveys (Guisan et al. 2006), projecting impacts of climate change (Pearson and Dawson 2003), predicting species invasions (Thuiller et al. 2005), to guiding species re-introduction (Pearce and Lindenmayer 1998), among others.

Due to the important role that SDMs play in research and as a tool for conservation of biodiversity, much effort has focused on the investigation of different technical aspects of SDM. For example,on a comparison of the performance of different modeling techniques in order to find the one most robust to different environmental conditions and characteristics of the occurrence data (Elith et al. 2006, Hernandez et al. 2006, Segurado and Araújo 2004, Farber and Kadmon 2003). Others have focused on the different statistical techniques that can be used to measure the performance or accuracy of predictions based on SDMs

(Fielding and Bell 1997, Liu et al. 2005). Also, on the theoretical and practical issues of the sources of uncertainties considering the different steps of building SDMs (Elith et al. 2002, Dormann et al. 2008).

Bias in the geographical location (i.e. spatial bias) of species records is a very important issue and a well recognized source of error in SDMs (Kadmon et al. 2004, Reddy and Dávalos 2003). It is rare to find a research paper where the authors do not acknowledge that one possible source of error of their predictions might be due to the geographical bias of species occurrences. However, investigation of the causes and ways to deal with spatial bias in the data has been a poorly endeavor in biogeography and macroecology and the few existing examples have clearly demonstrated how prediction accuracy is negatively affected when using biased data (Kadmon et al. 2004, Phillips et al. 2009). In this chapter we introduce a methodology and show an implementation to deal with spatially biased data.

SDMs are built based on the set of recorded geographical locations where the target species has been found. This information is normally obtained directly from field work campaigns, but also from museum or herbarium collections as well as from atlases and monographs. Nowadays, another important and frequently used source of information are freely-available biological databases that can be accessed and queried on-line (e.g. global biodiversity information facility). Most probably the data obtained from any of the sources mentioned above will come with a high degree of spatial bias. This bias is a direct representation of survey effort, meaning that some places were more frequently sampled due to accessibility (e.g., close to roads or close to cities) or because they places of species research focus (e.g. protected areas).

One important characteristic of the biological databases use in SDMs and in application at local to global extents is that the data consist of presence-only information. Traditional and modern modeling techniques (e.g. Generalized Linear Models, Generalized Additive Models, Boosted Regression Trees) rely on presence and absence data. Some of them are even robust against spatially biased data (see figure 1b in Phillips et al. 2009). However, to deal with presence-only data a new set of modeling methods has been recently developed (e.g. 'Ecological Niche Factor Analysis ENFA'; (Hirzel et al. 2002), "genetic algorithm for rule-set prediction"; (Stockwell and Peters 1999), "maximum entropy"; (Phillips et al. 2006)). Nonetheless, presence-only modeling techniques make use of a set of samples of the available environment to reach a better discrimination of the areas suitable and not suitable for any target species. These samples are known in the literature as pseudo-absences or background data. Here we will call them background data.

The approach commonly implemented to generate background data is to select a random sample of the background information (Elith et al. 2006). However, model prediction will be affected by combining random background information of the available environment

and spatially biased occurrence data (Phillips et al. 2009). The latter statement would be wrong if the spatial bias of occurrence data does not represent environmental bias. For example, collection records might be all situated near roads but if the network of roads properly represents the environmental variability in the study area then the performance of environmentally-based models will not be affected (Kadmon et al. 2004, Phillips et al. 2009). As this is not the case for the database of vascular plants for West Africa presented in the previous chapter, where there is rather a strong environmental bias in the location of the collection localities, it is then necessary to implement a methodology that allows presence-only methods deal with spatially biased data.

Phillips et al. (2009) proposed an approach where the selection of background data should reflect the same sample selection bias as the occurrence data. Models derived from the combination of both data sets will uncover any differences between the environmental condition where a species occurred and the environmental condition of the background information rather than reflecting patterns of sampling effort (Phillips et al. 2009).

One of the constraints of manipulating the background data in order to produce a subset having the same geographical bias as the presence data is that the sample selection distribution of the target species is normally unknown. To overcome this problem Phillips et al. (2009) limit the selection of background data for any species (e.g. one plant species) to the sites containing all records of the same species group (e.g. all plant species) assuming that all species in that group have been collected following the same scheme and sampling strategy. Their rationale is that the whole set of records for any group reflects survey effort (i.e. sampling bias). They called their approach 'target-background'.

We hypothesize that the database evaluation procedure presented in the previous chapter, specifically the resulting gap selection index, which adequately represent the spatial distribution of the sample selection in the database, can be used as a weight criterion for the selection of background data (thereafter called 'index background') and that model performance will be significantly better than using random or target background sets.

The first goal of this study is to make a comparison of the performance of different modeling techniques using three different sets of background information (i.e., random background, target background and index background). The target group are all vascular plants in the database of West Africa presented in the previous chapter. The second goal is to qualitatively evaluate for bias correction in vascular plant richness patterns as estimated by super-imposing individual species predictions based on the three techniques mentioned above. We expect richness patterns to match the known north-south richness gradient in the study area if models were corrected for spatial bias in the collection localities.

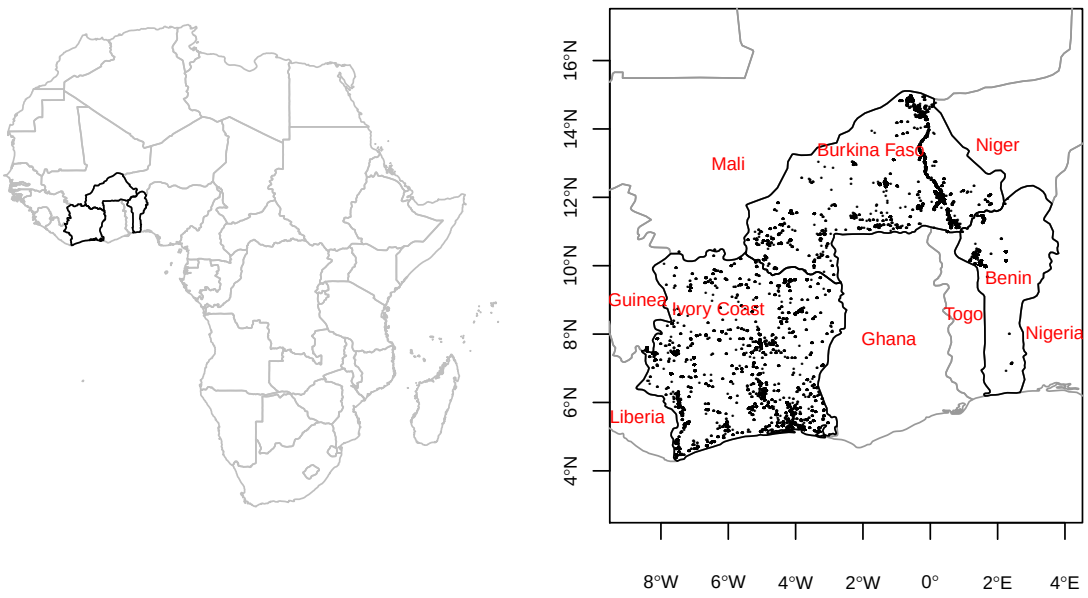

Figure 3.1 – *Left panel: Map of Africa showing in red the countries forming the study area. Right panel: detailed map of the study area; black dots represent collection localities*

3.3 Materials and Methods

3.3.1 Study area

The study area encompasses 730,600 km^2 in the countries included as part of the BIOTA project transects in West Africa (i.e. Ivory Coast, Burkina Faso and Benin) (Figure 3.1). The terrain is generally flat with a mean elevation of 277 meters above see level. However, some mountainous areas at the west side of Ivory Coast reach an altitude of 1500 meters above see level.

The study area is characterized by a continuous climatic North/South-west gradient. Annual mean temperature ranges from 29.6 degrees Celsius in the northernmost part of the study area in the Sahelian region to around 18.8 degrees Celsius in the western part of Ivory Coast. Total annual precipitation shows the opposite gradient: it ranges from 300 millimeter per year in the northernmost regions to more than 2,600 millimeter per year south west.

3.3.2 Vascular Plants Database

The database used in this study was originated from the compilation of different sources. The database of vascular plants in Burkina Faso and Benin has been described in Schmidt et al. (2005). This database is mainly the result of extensive field campaigns done in the

study area by researches and students. The database covering Ivory Coast was originated in the Botanical Garden of Geneva as part of the GIS-Ivory project (Chatelain et al. 2001). All databases were joined, corrected taxonomically, and filtered in order to extract records with a minimal spatial accuracy of 10 km^2. Finally, only species with 10 or more occurrences were selected for analysis. This number has been recommended in previous studies as the minimum sample size with which prediction performance is still good (Stockwell and Peterson 2002).

The final database consists of a total of 41,533 single observations distributed over 2,755 collection localities (Figure 3.1). Collection localities refer to relevés, geo-referenced herbarium collections and geo-referenced localities extracted from atlases and different gazetteer sources. The mean number of observation per collection locality is 15 with one location having 643 observations and 476 locations having only one observation. There are a total of 1,423 species belonging to 718 genera and 133 families.

3.3.3 Environmental Data

A set of 29 environmental variables was prepared to use as proxy to model the distribution of all species in the database (Table 3.1). All variables were resampled to match the spatial resolution of the vascular plants database (i.e., 10 km^2). Categorical variables (i.e. biogeographical zones and land cover) were resampled by selecting the value that occurred most frequently in all cells within 10 km^2 (i.e., the mode). Continuous variables were resampled by calculating the mean of all cells within 10 km^2.

Since many of the variables presented high levels of correlation with each other (Figure 3.2), we run a principal component analysis (PCA) on all variables. Categorical variables were split into single binary layers each layers representing each of the categories. Then, we selected the number of PCA axes that explained 80% of the cumulative variance (i.e., 6 components). As a result, the first six PCA components were used as predictor variables. In preliminary analysis we selected the number of PCA axes (i.e., 9) that explained 90% of the cumulative variance. However, the last 4 PCA components had recursively a significantly small influence on model fit and species predictions.

3.3.4 Background treatment

Three different sets of background data were used.

The first one consists of 5000 randomly chosen sites from the study area (referred to as 'random background'). This approach is the most common strategy employed in applications of models using occurrence and background data (Elith et al. 2006).

Table 3.1 – *List of environmental variables use for analyzes. The variable aspect ('aspect') was recalculated by taking the sine of the raw aspect direction on the 0 - 360° scale giving values between 0 and 1. Biogeographical zones were defined as: 1 = Sahel Zone, 2 = Sudanian Zone, 3 = Guinea/Congolia-Sudanian Zone, 4 = Guineo/Congolian Zone. Land cover categories were grouped in 5 classes: class 1 = categories 1,2,3, class 2 = categories 6,7, class 3 = categories 10,11, class 4 = categories 13 to 17 and class 5 = categories 18,19,20 (categories refer to the original classification of the land cover 2000). * Categorical variable*

Description	Abbreviation	Source
Annual Mean Temperature	temp	Hijmans et al. (2005)
Mean Diurnal Range (Mean of monthly (max temp - min temp))	temp_diur	Hijmans et al. (2005)
Temperature Annual Range (P5-P6)	temp_range	Hijmans et al. (2005)
Isothermality (P2/P7) (* 100)	isot	Hijmans et al. (2005)
Standard deviation of the mean monthly maximum temperature	tmax_std	derived
Standard deviation of the mean monthly minimum temperature	tmin_std	derived
Annual Precipitation	prec	Hijmans et al. (2005)
Precipitation of Driest Month	prec_drym	Hijmans et al. (2005)
Precipitation Seasonality (Coefficient of Variation)	prec_sea	Hijmans et al. (2005)
Precipitation of Driest Quarter	prec_dryq	Hijmans et al. (2005)
Standard deviation of the 12 monthly precipitation data	prec_std	derived
Elevation	elev	Hijmans et al. (2005)
Variance of elevation values (SRTM30) within a 9x9 moving window	elev_var	derived from elevation
Slope in degrees	slope	derived from elevation
Aspect	aspect	derived from elevation
Wetness Index	weti	derived from elevation
Channel Network Base Level	chan	derived from elevation
Proximity to water bodies	prox_wat	derived from elevation
Biogeographical zones	biog(1-4)*	White (1983)
Land Cover - GLC2000	lan(1-5)*	Bartholomé and Belward (2005)
Percent of bare ground cover	bare	Hansen et al. (2003)
Percent of herbacious ground cover	herb	Hansen et al. (2003)
Percentage of tree ground cover	tree	Hansen et al. (2003)
Annual average of spectral response values in the Near-Infrared, band2	spec2	derived from SPOT composites
Annual average of spectral response values in the Red channel. Band3	spec3	derived from SPOT composites

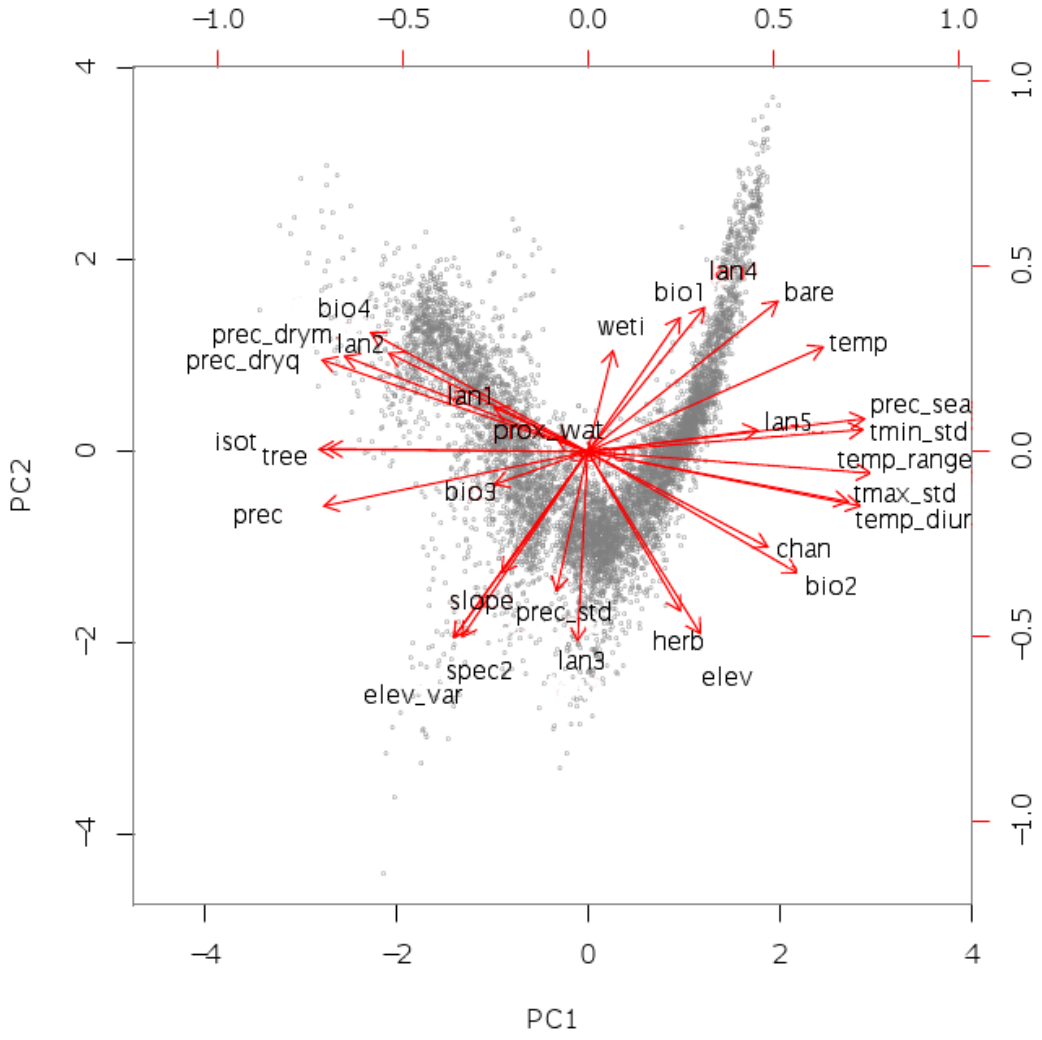

Figure 3.2 – *Biplot showing the relation between the environmental variables (red vectors) and the estimated values for all observation (i.e. all pixels) for the first two PC axis of the principal component analysis (gray dots). Two clusters of variables can be identified at the left and right sides of the plot. Variables within each cluster are highly positive correlated and variables in each of the two clusters are highly negative correlated. All variables in these two clusters load high in the first axis of the PCA. On the other hand, there is a cluster of variables in south direction. The variables forming this cluster load high on the second axis of the PCA and have low correlations with the variables in the first two clusters.*

In the second test, and as done in Phillips et al. (2009), background sites were selected for each species individually consisting of the presence localities of all other species (referred to as 'target background').

The third set of background data was generated by selecting random points in the study area but by specifying a probability density function. The values of this function were taken from the 'gap selection index' developed in the previous chapter. Values of the index range from 0, representing well investigated areas, to 1, representing areas in need of more information (i.e., negative bias). The probability density function was then defined as the inverse of the gap selection index so that the density of background points is higher in well investigated areas (referred to as 'index background').

For the first and third sets of background data we made sure to remove sites that by chance had the same coordinates as presence sites.

3.3.5 Modeling species distribution and estimating richness patterns

Several different techniques and algorithms exist that make use of presence and background information. We used the maximum entropy approach, implemented in the stand alone and free software Maxent, Version 3.3.1 (Phillips et al. 2006). In comparison to other modeling techniques, Maxent has constantly ranked within the best models (e.g. Elith et al. 2006, Hernandez et al. 2006) and has successfully been used to model species with as few as 5 occurrence records (Pearson et al. 2007). For a detailed explanation of the algorithms behind the techniques and its applications to species distribution modeling see Phillips et al. (2006; 2004). We run Maxent with default values for all parameters, except for the background data, which we independently prepared as explained in the previous section.

We transformed continuous probability predictions for all species to binary predictions, where 1 represents the occurrence range of the species and 0 areas not suitable for the species. As a criterion for this transformation we used the threshold given by Maxent and defined as the value where the proportion of correct classified cases (i.e. sensitivity) equals the proportion of wrong classified cases (i.e., specificity) of the training data.

Species richness was then estimated by overlapping and summing binary predictions. For this purpose, we used only those species that were accurately predicted, that is, those species that achieved an AUC equal or greater than 0.8. (see section 3.3.6 for an explanation of the AUC). From a comparison of the richness patterns generated by using species modeled with different background approaches, we did a qualitative evaluation of the influence of approaches that seek to contra-rest the bias influence in model predictions (target and index background).

3.3.6 Model performance evaluation

For each species, occurrence locations were joined with the three types of background data described in section 3.3.4. Then, occurrence-background sets were split into training and testing sets keeping always a proportion of three to one respectively and at the same time preserving relative ratios of presence and background sites.

To evaluate the accuracy of predictions we calculated the area under the curve (AUC) of the receiver operating characteristic (ROC) technique. The ROC is a threshold-independent technique that compares the fraction of correct classified cases (i.e., *sensitivity*) against the fraction of wrong classified cases (i.e., $1 - specificity$) for all possible thresholds (Fielding and Bell 1997). It measures the probability, that in a pair of randomly chosen out of the presence and absence data, the model will assign a higher probability of occurrence to the case with the observed presence (Bonn and Schröder 2001).

To test for statistically significant differences between model predictions for each species between the three background treatments we used the McNemar test (McNemar 1947). Since our training and testing sets had been generated from the same database, they are not independent. The McNemar test has demonstrated to be less sensitive to dependence in the data and has achieved better discrimination power than other typical model comparison techniques (de Leeuw et al. 2006). For calculation of the McNemar test we compared testing sets with binary predictions for each species independently.

3.4 Results and Discussion

An obligatory step when modeling the geographical distribution of several species is to measure predictions performance. This is especially relevant if the objective is to estimate species richness patterns based on the cumulative sum of species' independent estimates. For the modeled species selected from our database a clear pattern is found where high model accuracy stabilizes as the number of occurrence localities increases (Figure 3.3). Only 13% of the modeled species (i.e., 189 species) scored an AUC of less than 0.8.

The general pattern of model accuracy decline with decreasing sample size has been found in several studies that have investigated this relationship (Hernandez et al. 2006, Stockwell and Peterson 2002). Stockwell and Peterson (2002) concluded that none of the methods used predicted consistently good with sample sizes less than 30 occurrences. This seems also to be valid for our findings. This consistency is lost with sample sizes less than 30 although good model performance was also obtained when sample size was lower than 30. Maxent has been ranked as one of the best models producing accurate predictions (Elith et al. 2006) and a modeling method capable to deal with sample sizes as small as 5 or 10 occurrences (Hernandez et al. 2006, Pearson et al. 2007), results that are corroborated

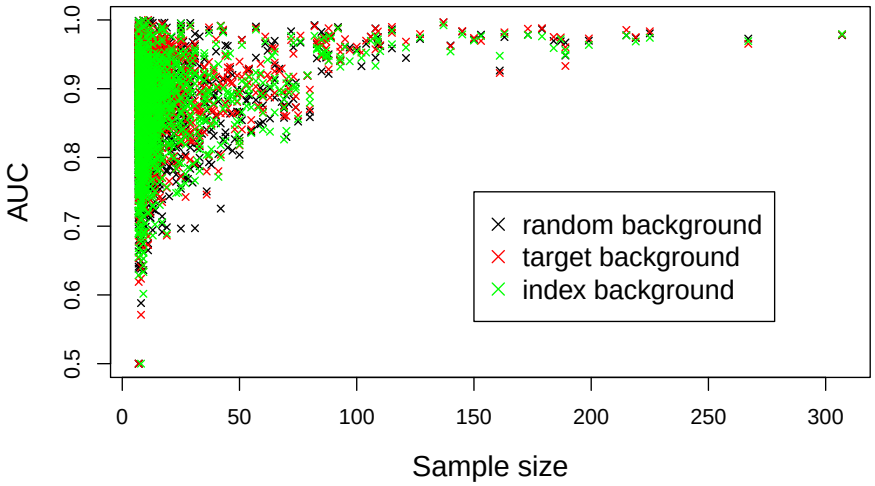

Figure 3.3 – *Relation between model performance and sample size. Model performance was measured with the area under the curve (AUC) of the receiver operating characteristic (ROC). Sample size refers to the number of occurrences used in model's training sets. The minimum sample size is 7 (i.e., 290 species) and the maximum 307 (one species, i.e.,* Zornia glochidiata*). 1,423 species in total were analyzed.*

in our study. Investigating the reasons why some species with sample sizes less than 30 occurrences were in some cases accurately predicted and in some others not is beyond the aims of this study.

When using random background the average AUC value was 0.88. The average AUC value slightly improved by using target background (i.e., 0.9). In contrast, the mean AUC value decreased when using index background (i.e. 0.87)(Figure 3.4). Despite of the small differences in AUC values for the three modeling approaches, there were in all cases a significant difference in model predictions in more than 90% of the cases ($P < 0.05$, McNemar test with continuity correction, paired by species)(Table 3.2).

In order to identify if model predictions have been corrected for bias in the input data, testing data sets that are not biased are needed. In our case, we do not have such testing sets and therefore we assumed that by using target and index background treatments we are correcting for bias in model predictions. Having that in mind, we can conclude that in agreement with Phillips et al. (2009) results, target background treatment not only corrected for bias predictions but also improved, in 91.5% of the cases, model accuracy.

Surprisingly, model performance decreases when using index background in comparison to both target and random background, although we assume that predictions were corrected for bias. Similar approaches to index background exist in the literature. For example, Zaniewski et al. (2002) created a model of survey effort and used it as weights for selection of background data. They found that model predictions were more similar to results of models based on real presence-absence data, than models relying on presence-random

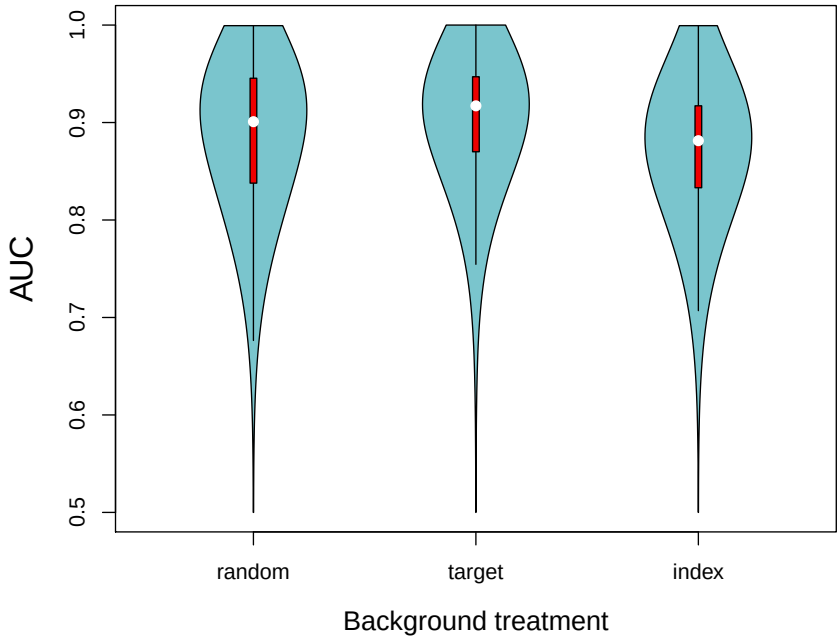

Figure 3.4 – *Comparison of model performance using three different background data sets on the test data sets, measured using the area under the curve (AUC) of the receiver operating characteristic. Blue areas represent the density of AUC values. White circles are the median and red bars are the values within the second and third quartiles.*

Table 3.2 – *Pair-wise comparison of model predictions based on different background treatments, measured using the McNemar test with continuity correction and paired by species. The table shows the number of cases (i.e. number of species, below the diagonal) and percentage (above the diagonal) of cases where there was a significant difference between model performance ($P < 0.05$). A total of 1423 species predictions were compared.*

background treatment	random	target	index
random	–	90.6%	91.7%
target	1290	–	91.5%
index	1305	1297	–

background and presence only models. Engler et al. (2004) created also weighted background data for each species independently based on a prior modeled distribution range of the species based on a presence-only model (ENFA). Background data were chosen in areas unlikely suitable for the species. In that case weights to select background data were given in the opposite direction as the one applied in this study and therefore their approach does not deal with bias in the data.

We evaluated if the magnitude of differences between AUC values of target background based models with random background based models have a particular relationship with sample size. In fact, AUC differences are constantly minimal when the sample size is greater than 100 occurrences approximately (Figure 3.5). Below this number there is not a clear relationship between sample size and AUC differences, with cases spread in areas

of small and big differences in AUC regardless of the sample size. This results add to the known fact of prediction uncertainty when small sample sizes are used to generate predictive models.It is also important to point out than only in 65% of the cases the difference in AUC was greater than 0, which means that predictions of 927 species (from the pool of 1,423 species) improved when using target background.

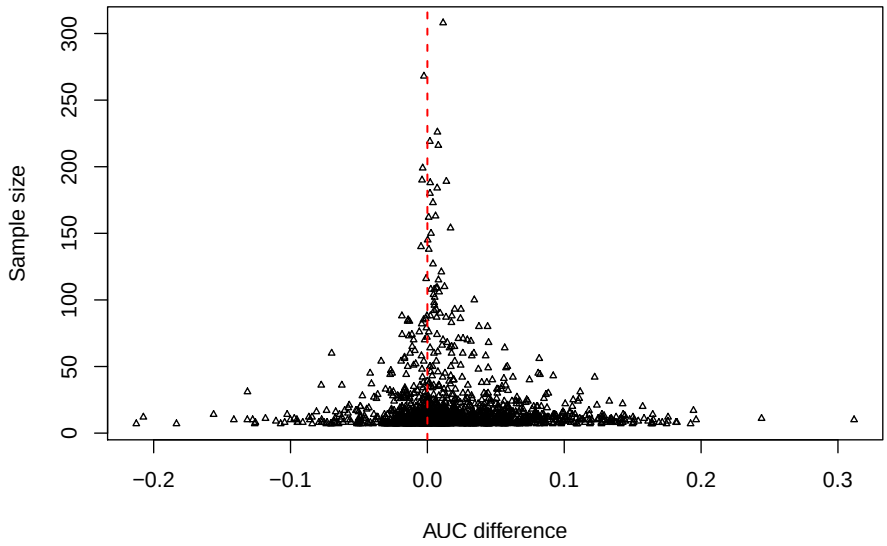

Figure 3.5 – *Relation between sample size (y axis) and the difference between the prediction accuracy values, as measure with the area under the curve (AUC) of the receiver operating characteristic, of models treated with target background and random background sets.*

We created prediction maps of three species having the biggest AUC difference between models based on target and random background in order to visualize prediction differences (Figure 3.6). The prediction models of the species *Cyperus dilatatus* achieved the greatest improvement from an AUC of 0.59 when using random background to an AUC of 0.9 when using target background. Evidently, the occurrence probability prediction of *Cyperus dilatatus* using random background is not better than a pure random prediction given the low AUC. Visually, there is no discrimination power at all (Figure 3.6, first panel in first row). On the contrary, by using target background and index background, there was a significant improvement on model performance and probability of occurrence values were better discriminated.

In general, models based on target background predict areas of high occurrence probability restricted to areas where sample locations are present (red points in figure 3.6). While areas apart from occurrence locations are predicted with low occurrence probability based on the target background approach, the same areas achieved higher probabilities values with the other two approaches. For example, although *Striga macrantha* samples are restricted to the Sudanian zone and below, models based on random background assigned this species greater occurrence probability values in the Sahelian zone, while models based on target and index background restrict predictions to the Sudanian zone.

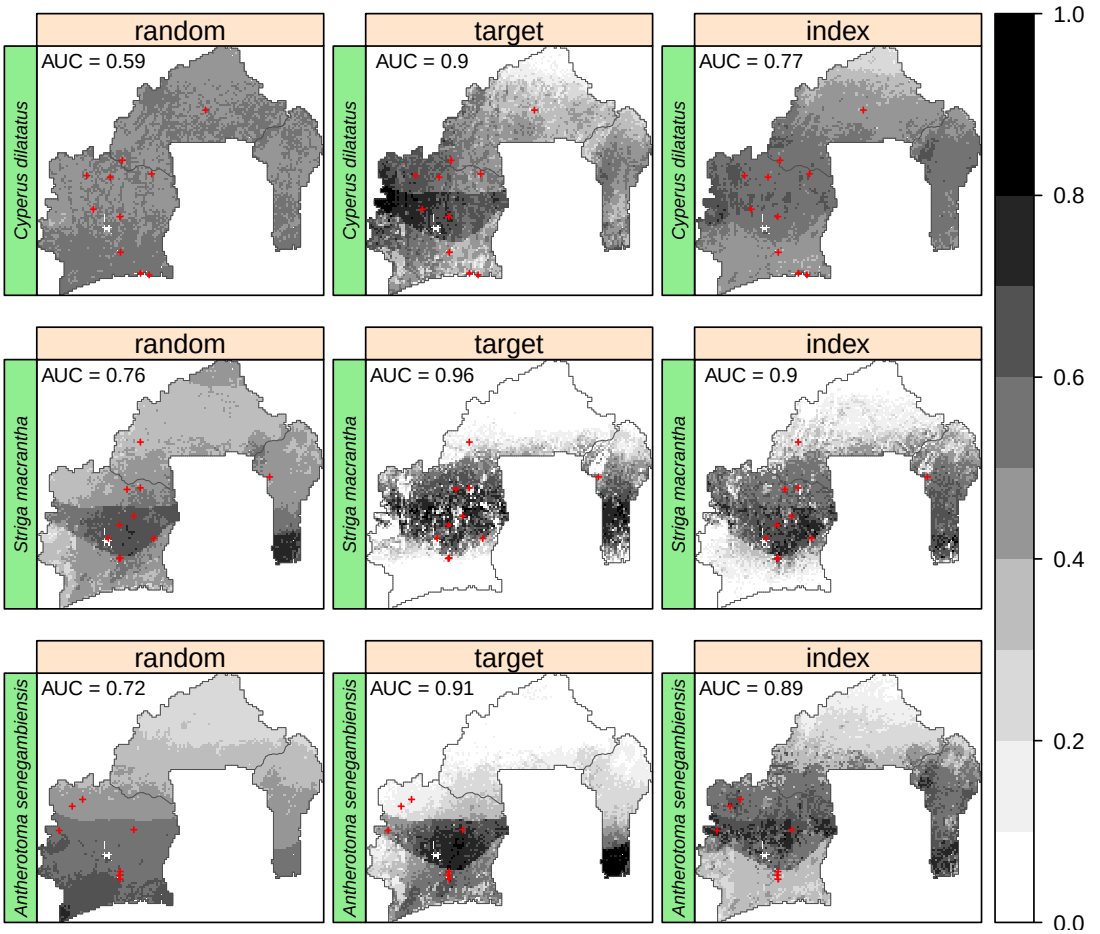

Figure 3.6 – *Occurrence probability maps of three species having the biggest difference between the AUC of models using target and random background sets. Also displayed and for comparison purposes are predictions based on models created with index background sets. Depicted in red are the sample locations used to train the models.*

By overlapping and adding binary maps of selected species (i.e., those with AUC values $\geq$ 0.8) we estimated richness patterns of vascular plants. Three different maps were obtained according to the background treatment predictions were based on (Figure 4.2). A visual checking of the maps allows us to qualitatively evaluate the influence of using biased data on resultant patterns. A well known diversity gradient in the study region exists with species richness increasing in a north-south direction. This gradient is accurately represented in the map based on target background. However, maps based on random and index background are far apart from this gradient and areas where species richness is higher than their surroundings seem to be affected by collection density (see figure 2.4 in chapter 2). For example, species richness in the Sahelian zone in the map created with predictions based on random background (Figure 4.2a) is higher than species richness in more southern areas in the Sudanian zone. The same occurs in north Benin, where maps a and c in figure 4.2 predict higher species richness than areas in the south. Again, those errors seem to be related to the density of collection localities located in those areas (see

figure 2.4 in chapter 2)

3.5 Conclusions

We have modeled the occurrence probability of 1,423 vascular plants in West Africa using one modeling technique (i.e., Maxent) and three variations of the background data (i.e., random, target and index background) required as input to this model. Although average model performance was high regardless of the background type used, by using index background we obtained in average better model performance and were able to correct for the spatial bias present in the collection localities. Bias was not successfully corrected for when using random and index background.

Even though collection records are spatially biased, biased predictions estimates can be avoided if sample size is big enough. We found minimal differences between model performance using random and target background when sample size was greater than 100 records (Figure 3.5). One of the priorities for future research is further collection of those species that at present are underrepresented in the database.

In general, model performance was significantly improved, in most of the cases, by using target background. Model predictions were improved by correcting for spatial bias when using target background. Individual species predictions using target background were more conservative by limiting prediction of suitable areas to those areas where collection records exist. Species richness patterns based on predictions of models using target background were more accurate to the known richness gradient patterns in West Africa. We therefore recommend the use of target background when the purpose of the study is to model species ranges and richness patterns using biased data.

We did also obtain good model performance when using random background. But one issue we have to consider is spatial dependence (i.e., autocorrelation) of the data we used to train and test the models. They are not independent because they originate from the same source, but techniques to measure model performance (e.g., AUC) generally assume independence between training and testing data sets. The use of non-independent data to build and test species niche models may generate an optimistic assessment of model performance (Araújo and Guisan 2006). Recently, Veloz (2009) also demonstrated how measures of model performance are inflated when the pattern of collection localities are clustered. That is, when they are spatially autocorrelated, which is the case for our database.

One of the main constraints of our database for its use in macroecological and biogeography studies (e.g., modelling species distribution) is the high amount of species with few records. By applying niche models to such data, prediction uncertainties increase (i.e., spatial

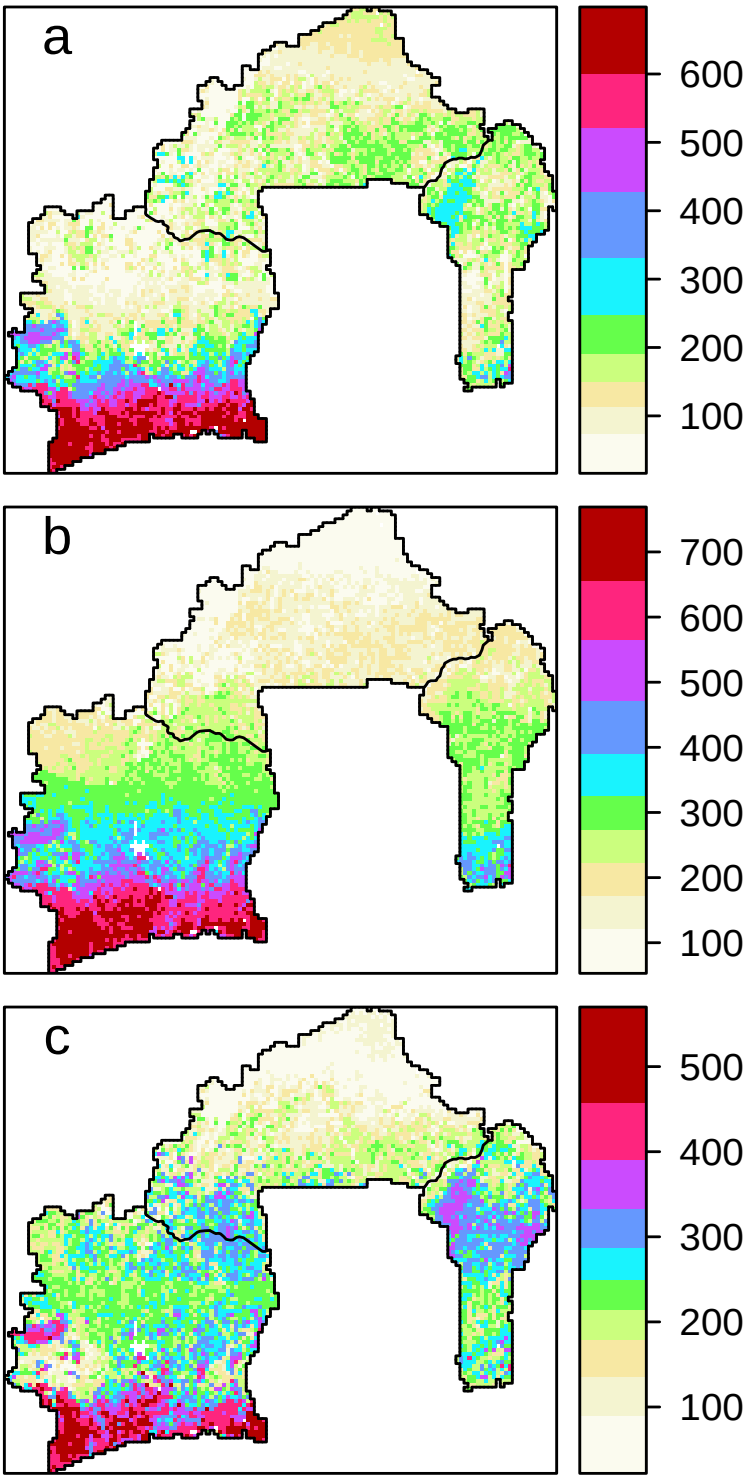

Figure 3.7 – *Patterns of vascular plants richness estimated using the maximum entropy technique (Maxent) using occurrence locations and three different background treatment types: random background (a), target background (b) and index background (c). Each species was modeled independently and only those with an AUC equal or greater than 0.8 were chosen to calculate species richness; 1192 species were selected using random background, 1310 species with target background and 1211 species with index background.*

pattern predictions and model performance are different from one model to another) and there is no rule of thumb for which model or model strategies to use. Methodologies to circumvent that problem exist in the literature. For example, ensemble forecasting, where predictions from different models are averaged out and uncertainties reduced by giving more weight to those areas commonly predicted as suitable in all independent predictions (Araújo and New 2006, Araújo et al. 2005).

Another approach suitable for our database is to include estimates of collection bias into the analysis of species ranges as exposed by Schulman et al. (2007). They made a topographical surface representing a landscape of collecting activity based on a network of Thiessen polygons created in relation to the location of collecting localities. Then, a circular buffer of a certain distance around each collection record was made and reshaped according to the surface of collection activity. The buffer did not change its form in the vicinity of areas where collection activity was high but it spread outwards in areas of few collection activity with the rationale that those areas are uncertain (because they have not been visited) and it is probable to find the species there. Their methodology however, requires a good knowledge of each species in order to generate the buffer (e.g., dispersion capabilities) and it can be applied to all species with few records.

Species distribution models are nowadays one of the main tools of analysis in macroecology and biogeography applications. With the increase in cost-free data and software availability, generating predictions of species ranges might be a very easy task. However, we have demonstrated how results of applying SDMs can result in misleading information. If such results will form the basis for decision making regarding the conservation of biodiversity, or the criteria to test ecological and evolutionary hypothesis, all the outcomes will be false. We recommend a critical use of SDMs by comparing different strategies in order to minimize prediction uncertainties and obtain the best possible results.

CHAPTER 4

CROSS-TAXON PATTERNS OF BIODIVERSITY IN WEST AFRICA

4.1 Abstract

The spatial congruence between distribution patterns of different taxa is a main criterion to prioritize conservation areas. Most important are those areas where high levels of diversity (i.e., "diversity hotspots") overlap. We used the most comprehensive databases of vascular plants, amphibians and bats in West Africa to evaluate the extent of spatial concordance between their distribution patterns. We applied a species-specific niche modeling approach (Maxent) to estimate species' geographical distribution ranges. By superimposing these predictions, species richness patterns and the range size rarity index were calculated for each group. Using spatial auto-regressive techniques we estimated the main environmental determinants of the geographical variation in the distribution of each group. Spearman correlation coefficients were calculated to evaluate overall pair-wise congruence between the taxa. In addition, correlations were calculated at small geographical extents using a moving window approach. Hotspots, defined as the top 5% grid cells with the highest range size rarity values, were identified for each taxon and percentage overlap between hotspots was calculated. Finally, the threat imposed to hotspots was quantified by superimposing current land cover patterns and protected areas and calculating their coverage. Temperature and elevation heterogeneity were the main determinants of variation in species richness and range size rarity patterns of the three taxa. Overall correlations between pair-wise comparison of species richness and range size rarity were

very high (> 0.9) and statistically significant in all cases. However, small extent variation in congruence was detected with areas of weak and negative correlation. Hotspots for all taxa together occupy 9.3% (i.e. 583.7 km^2) of the study area of which 39.2% is still covered by natural vegetation and 8.5% covered by the network of protected areas. The fact that only small areas of congruence of high species richness and endemism between the three groups were identified in this study, and that they are poorly covered by the network of protected areas, represents a call of attention for a more efficient planning of conservation efforts in West Africa.

4.2 Introduction

Defining areas for biodiversity conservation could be an easy task if there are similarities in the geographical distribution of different taxa. In fact, actual distribution patterns of species can be thought as the result of a long history of interaction between different taxa, suggesting the potential for similarities in their distribution (Jetz et al. 2008). There is a set of ecological hypothesis trying to explain the causes and origin of past and current biodiversity distribution patterns. If in fact, actual distribution patterns can be accurately explained by this set of ecological hypotheses, then a high degree of congruence between different taxa is also expected. Some hypothesis state that an increase in niche heterogeneity and availability promotes species diversification and coexistence, either because of more resources available (Hutchinson 1959, Chesson 2000) or different degrees of vegetation structure complexity (Kissling et al. 2008). Other hypotheses stress the influence of indirect factors related to the environment, like climate, topography, and habitat heterogeneity on actual species distribution patterns(Currie et al. 2004, Hawkins et al. 2003, Kreft and Jetz 2007).

Generalizations about positive or negative spatial congruence of different taxa can not be easily made. Congruence patterns vary between different taxa, from one region to another and between spatial scales. For example, Prendergast et al. (1993) investigated the congruence of five taxa in Great Britain and found that areas rich in species of one taxa generally do not coincide with other taxa, while Jetz et al. (2008) found a strong positive correlation between producer and consumer diversity on a global scale. The same conclusion was reached by Lamoreux et al. (2006) who showed that global patterns of richness and endemism are highly correlated between different taxa. Conclusions about congruence of cross-taxon analysis depend on both, scale (i. e. grain size) and extent of analysis (Pearson and Carroll 1999).

The effectiveness of natural resources management, and specifically biodiversity, strongly depends on specific measures employed to define target sites for conservation. Species

richness (i.e., the number of species in a given area) and endemism (i.e., number of range-restricted species) are two of the criteria commonly used to define priority areas for conservation (Myers et al. 2000). But even then, species-rich and endemism-rich areas do not necessarily coincide within a single taxonomic class, as exemplified by bird species on a global scale (Orme et al. 2005). A combined index (i.e., range size rarity), that puts together these two criteria (i.e., species richness and endemism) has been developed by Williams et al. (1996) and has undergone further refinements since (Kier and Barthlott 2001). The use of this index has improved the selection of areas for the conservation of diversity by increasing the number of species represented in target areas (Williams et al. 1996). In this study we selected this index to evaluate patterns of congruence between different taxa as well as to define important areas of diversity (i.e., range size rarity hotspots) for three different groups (i.e., vascular plants, amphibians, and bats).

The Upper Guinean Forest of West Africa has long been recognized as one of the world's hotspots of plant diversity (Myers et al. 2000, Küper et al. 2004a). It is also an area acknowledged for its large number of endemism (Brooks et al. 2001). The focus of research in this area has been on the floristic composition, the ecosystems services and threats to the forest (e.g. Poorter et al. 2004a, Hawthorne and Jongkind 2006). It is estimated that the Upper Guinean Forest contains 2800 plant species of which $\approx 23\%$ are endemic (Poorter et al. 2004b). Deforestation, habitat fragmentation, and over-exploitation, among others, are some of the threats imposed to the vegetation in this region, and it has been speculated that the disappearing of the vegetation directly affects animal populations. Still, little is known about geographical patterns of animal species richness in the region. If the spatial distribution patterns of plant and animal richness and endemism coincide in this area, then specific remnants can be defined that contribute to the conservation and persistence of biodiversity as a whole.

For this study we used the most comprehensive available databases of vascular plants, amphibians and bats in West Africa. The first aim of this study was to estimate species richness and the range size rarity index for each taxa. The second aim was to evaluate spatial congruence patterns between species richness of the three taxa, as well as between the range size rarity. Congruence patterns are also evaluated for hotspots defined for each group. The final aim was to quantify threats imposed to hotspots areas by current land use.

4.3 Materials and Methods

4.3.1 Species Databases

Geo-referenced location records of vascular plants, amphibians and bats were compiled for several countries in west Africa (Figure 4.1). Data sources are very heterogeneous, ranging

from field work collections, herbarium vouchers, geo-referenced maps, among others. Data for vascular plants has been extracted from the Biogeographical Information System on Africa Plant Diversity (BISAP) database (see Küper et al. 2006, for a detailed description).

Figure 4.1 – *Map of the study area in West Africa. The study area is represented by countries colored in gray.*

To prevent model performance problems (Pearson et al. 2007), only records containing 3 or more observations per species were considered for analysis. In total, 752 vascular plant species, 158 amphibian species and 110 bat species were modeled. Due to the heterogeneous sources of information, the geographical location of all species was standardized to a minimal common resolution of 0.5 degrees ($\approx 50\,\mathrm{km}^2$).

4.3.2 Environmental Layers

A set of environmental variables was prepared to model the distribution range of vascular plants, amphibians, and bats (Table 4.1). Variables were resampled to a 0.5 degree cell from their original resolution to match the resolution of the species databases. The three categorical variables (land cover, soils and WWF ecoregions) were resampled by selecting the value that occurred most frequently in all cells within each half degree cell (i.e., the mode). All other variables, which have continuous values, were resampled by calculating the mean of all cells within each half degree cell.

We checked for strong collinearity between environmental variables. Removing collinear variables is important for accurate coefficient estimation (Graham 2003)(see subsection Determinants of species richness and endemism). We removed all variables with a correlation coefficient greater than 0.6. Five variables remained for analysis (i.e., prec_min,

Table 4.1 – *List of environmental layers used to model the distribution of vascular plants, amphibians and bats in West Africa. *Variables used to model vascular plants but not amphibians or bats*

Variable Name	Abbrev.	Source	Original Resolution
Percent of bare ground cover	bare	MODIS	500 m
Percent of herbaceous ground cover	herb	MODIS	500 m
Percent of tree ground cover	tree	MODIS	500 m
Annual average of spectral response values in the Near-Infrared channel, Band2	spec2	SPOT-VEGETATION composite	
Annual average of spectral response values in the Red channel, Band3	spec3	SPOT-VEGETATION composite	
Proximity to water bodies	prox_wat	DCW data	1 km
Maximum value ("wettest month") of the 12 monthly precipitation	prec_max	Worldclim1.4	1 km
Minimum value ("driest month") of the 12 monthly precipitation	prec_min	Worldclim1.4	1 km
Standard deviation of the 12 monthly precipitation data	prec_std	Worldclim1.4	1 km
Total annual precipitation calculated as the sum of all 12 monthly rainfall	prec	Worldclim1.4	1 km
Maximum of the mean monthly maximum temperature	tmax_max	Worldclim1.4	1 km
Minimum of the mean monthly maximum temperature	tmax_min	Worldclim1.4	1 km
Standard deviation of the mean monthly maximum temperature	tmax_std	Worldclim1.4	1 km
Maximum of the mean monthly minimum temperature	tmin_max	Worldclim1.4	1 km
Minimum of the mean monthly minimum temperature	tmin_min	Worldclim1.4	1 km
Standard deviation of the mean monthly minimum temperature	tmin_std	Worldclim1.4	1 km
Contrast (range: max-min) of elevation values within a 3x3 moving window	elev_con	NASA/glc	1 km
Variance of elevation values (SRTM30) within a 9x9 moving window	elev_var	NASA/glc	1 km
Global land cover 2000*	glc	Bartholomé and Belward (2005)	1 km
Soil*	soil	soil	1 km
WWF Ecoregions*	wwf	wwf	1 km

prec, elev_var, tmax_min and tmin_max). Variables prec, prec_min, and elev_var were transformed to the logarithm scale. All variables were standardized to zero mean and unit variance.

Both environmental variables, prec_min and prec describe the north-south precipitation gradient in the study area. However, the former emphasizes dry patterns while the latter wet patterns. The variable elev_var is an indicator of terrain variability which translates in habitat diversity.

4.3.3 Statistical Analysis

Modeling Species Richness and Endemisms

We used the maximum entropy method (as implemented in Maxent, (Phillips et al. 2006)) to predict the geographical distribution of all species. This method adapts well to our databases since we are dealing with presence-only data and with a high proportion of species containing few location records. Maxent has also proved to produce more accurate results when compared to other modeling techniques (Elith et al. 2006).

We modeled vascular plants, amphibians and bats using all location records and Maxent default parameter values. To convert continuous predictions of amphibians and bats to binary predictions we chose the equal test sensitivity and specificity threshold provided by Maxent. The resulting binary predictions were further refined by expert knowledge. For vascular plants we chose the highest threshold provided by Maxent in order to minimize prediction error and assure that areas predicted as suitable represented only those areas where the probability of occurrence was high.

Since our main goal was to obtain predictions and not to evaluate variable importance, we modeled all species using the complete set of environmental predictor without regard to collinearity issues.

Species richness was calculated by superimposing all binary predictions per grid cell. Endemism was calculated using the range size rarity algorithm (Williams et al. 1996). It is calculated for each grid cell and it considers the inverse range size of all species present in each particular cell, that is, the inverse of the sum of the number of grid cells occupied by each species (4.1) (Usher 1986).

$$\textit{Range Size Rarity Index} = \sum 1/C_i \qquad \{i : C_i \neq 0, 1 \leq i \leq n\} \tag{4.1}$$

where C_i is the number of grid cells occupied by species i.

Determinants of species richness and endemism

In a first step, all richness and endemism estimates were log-plus-1 transformed to make them follow a normal distribution. Secondly, determinants of geographical variation in species richness and endemism were calculated by estimating variable coefficients of spatial and non-spatial models. One aspect that can harm coefficient estimation is the spatial autocorrelation of both, response and predictor variables (Dormann 2007a).

To deal with this, we performed the analysis by comparing non-spatial models (i.e., Generalized Linear Models (GLM)) and three spatial models (i.e., Simultaneous Autoregressive Models lagged-response (SAR_lag), lagged-mixed (SAR_mix), and spatial error (SAR_err)). All SAR models were calculated using the same weight matrix with a neighborhood distance of approx. 100 km, i.e., a pixel unit distance and a coding style "W" (i. e. row standardized). All models are implemented in the free software R (R Development Core Team 2009) using the library spdep (Bivand et al. 2009). To test for spatial autocorrelation we performed Moran's I analysis on the residuals of all models and plotted it using correlograms. Spatial autocorrelation is absent if the Moran's I values approach 0 (Fortin and Dale 2005).

Cross-taxon Congruence

To evaluate the overall pair-wise congruence between species richness and endemism, we calculated the Spearman's rank correlation coefficient to accommodate the non-normal distribution of species richness and endemism values. For each correlation pair we run a standard significant test of the null hypothesis of the correlation value being zero.

To identify small scale (*sensu* extent) variation in pair-wise congruence of richness and endemism, we calculated the correlation coefficient for each 0.5 degrees cell using a moving window of $\approx$ 250 km size. We selected this particular distance because the number of neighborhood cells in the moving window was more or less constant throughout the study area, and because the influence of cells around the coast versus inland cells was not as big as when using greater distances. On the contrary, smaller distances resulted in too few grid cells within the moving windows to calculate correlation coefficients.

Hotspots definition and threat analysis

General correlation analysis do not make explicit which areas with high correlation are originated from the congruence of high species richness or high range size rarity values for both taxa compared (i.e., hotspots). The definition of hotspots was based on the values of the range size rarity index. First, all grid cells were ranked from high to low range size rarity. Secondly, the number of grid cells in the highest five percentile was defined as hotspot. Congruence between hotspots was done by calculating the amount of spatial overlap between them.

We used the global land cover dataset (GLC2000, Bartholomé and Belward 2005) provided by the European Joint Research Centre (JRC) to evaluate the threat of current land use to the hotspots. As a first step, we reclassified all categories in the original dataset into two new classes: Natural Areas and Intervened Areas. The first one is composed of original

classes 1 to 15 and the second of classes 16 to 23. In a second step, we calculated the coverage of natural and anthropogenically influenced areas inside the hotspots. We also used the current network of protected areas (World Conservation Union and UNEP-World Conservation Monitoring Centre 2007), to evaluate its coverage within hotspots.

4.4 Results

Richness patterns of the three groups followed the known north-south climatic gradient in West Africa, with few species in the north and large richness numbers near the coast (Figure 4.2). However, clear differences exist when patterns are analyzed in more detail. For example, while areas rich in bat species are situated in the ecotone between forest areas and savannas, the highest richness of vascular plants and amphibians occurs near the coast.

Areas of high range size rarity for vascular plants stretched out throughout the coast from Nigeria to Sierra Leone, except from Benin and Togo (Figure 4.3). For amphibians and bats areas of high range size rarity are geographically more restricted to Nigeria, next to the border with Cameroon (amphibians) and between Ivory Coast, Guinea, Sierra Leone and Liberia (bats).

Another important remark is that all bats together seem to occupy wider habitat ranges than amphibians and vascular plants given the lowest maximum value of the range size rarity index (i.e., 0.48). Species with more restricted ranges will increase the endemism index values calculated per grid cell as in the case for vascular plants and amphibians (i.e. maximum of 4.68 and 3.48 respectively)

Strong spatial autocorrelation is evident for the non-spatial model (GLM) when looking at the residuals correlogram plots (Figure 4.4). It is also explicit how all spatial autoregressive models do correct for spatial autocorrelation. The lowest AIC was achieved by the SAR_mix model and we chose it as our best model. This result was expected since the spatial autocorrelation of our data is present in both, species richness counts and in the predictor variables. Mixed models are known to deal with this issue (Haining 2003).

Many similarities were found regarding the environmental determinants of species richness and the range size rarity for all groups (Table 4.2).

A positive and significant relationship exists between species richness of the three groups and temperature (i.e., tmax_min) and the variance of elevation values (i.e., elev_var). Unexpectedly, total annual precipitation has a significant but negative influence on vascular plant richness. We expected to find areas rich in vascular plant in regions with the highest values of precipitation. One possible explanation is that there are very few samples of

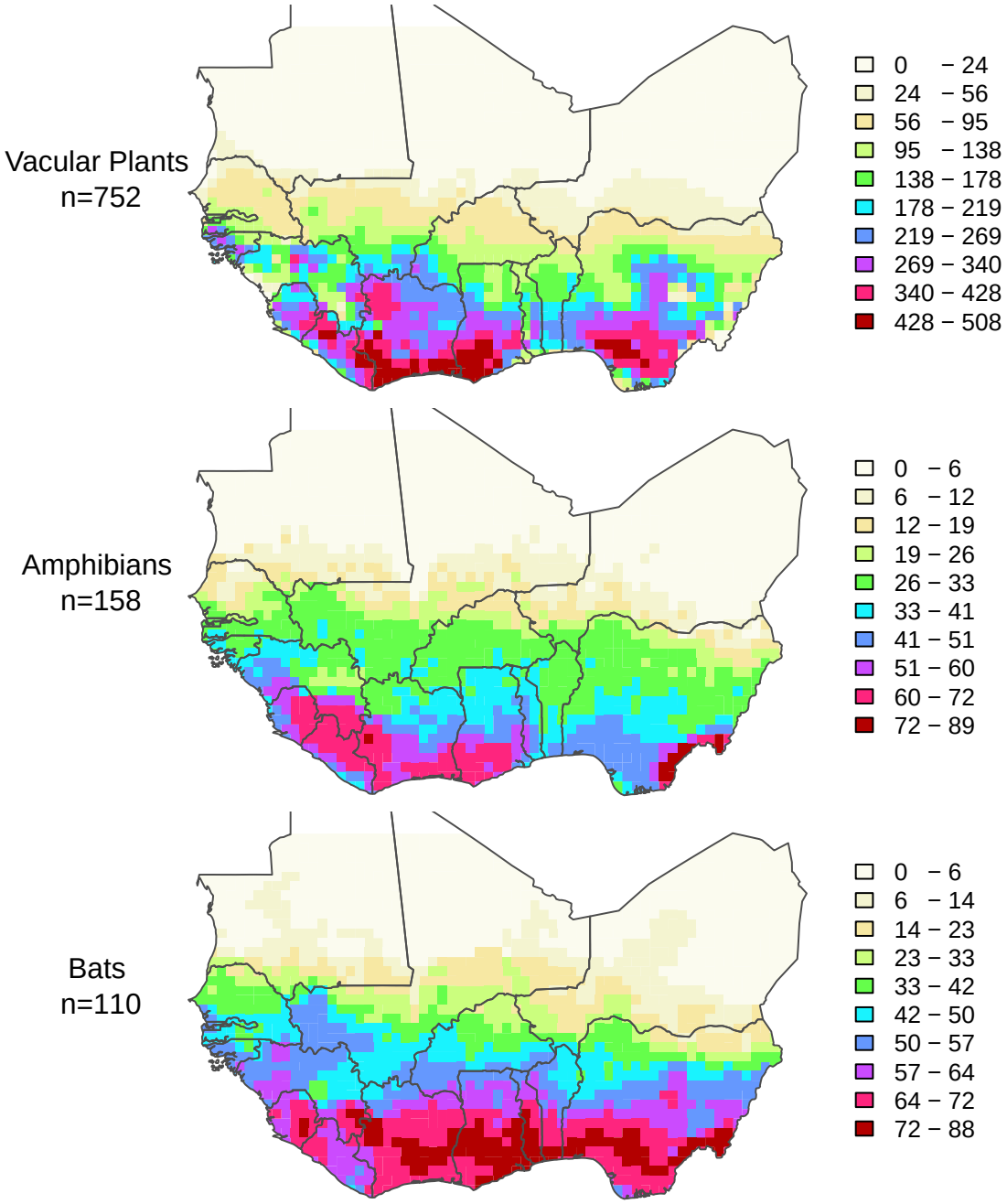

Figure 4.2 – *Geographic variation of species richness of vascular plants (n=752), amphibians (n=158), and bats (n=110). Species richness were calculated by superimposing prediction ranges of all species modeled*

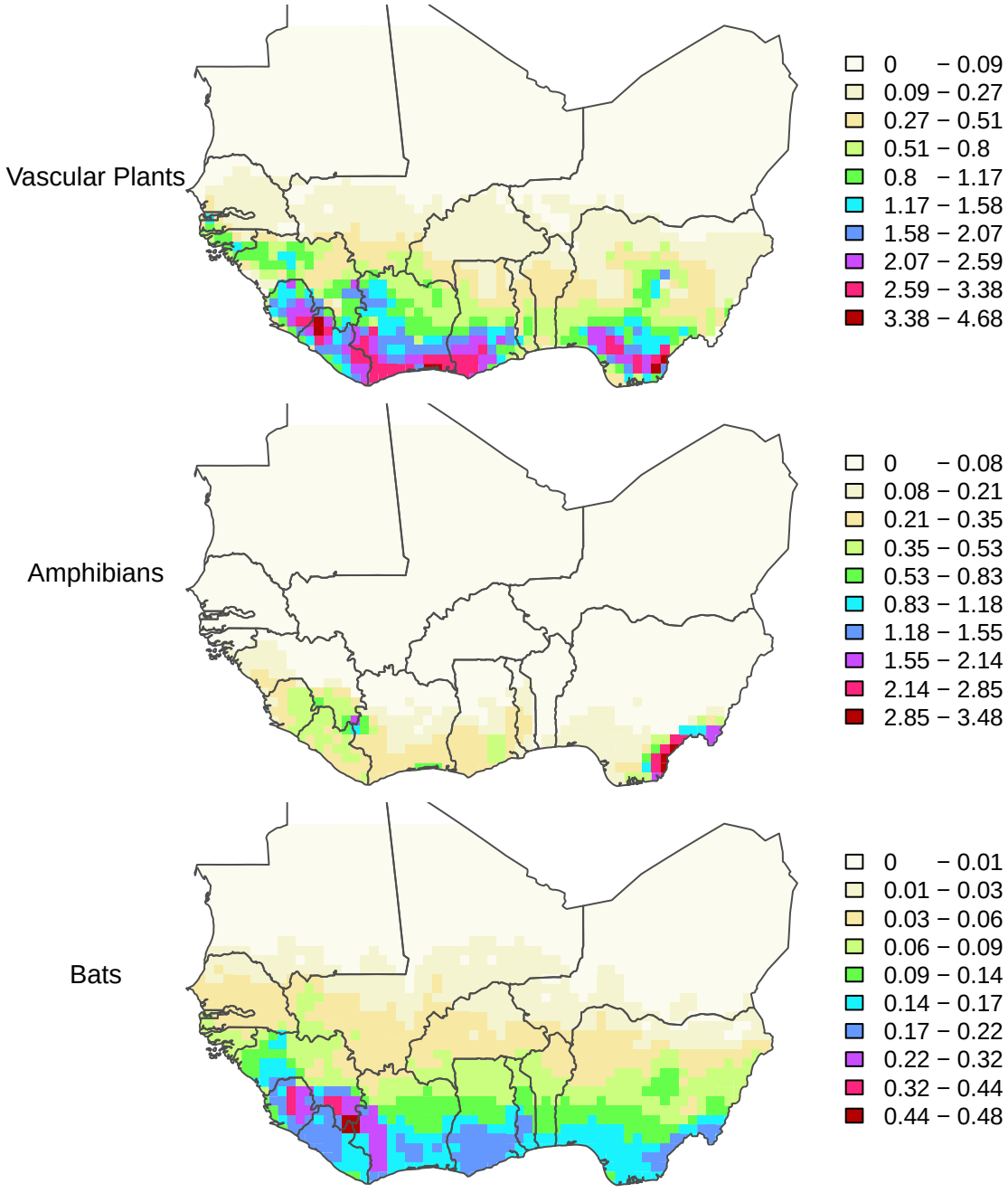

Figure 4.3 – *Geographic variation of the range size rarity index for vascular plants, amphibians and bats.*

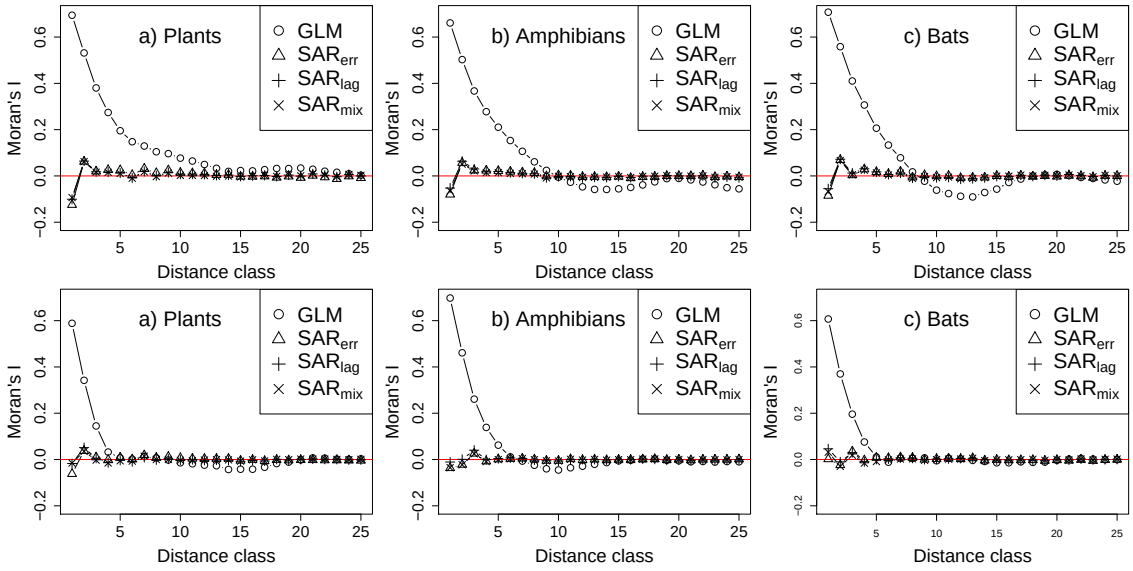

Figure 4.4 – *Correlograms for residuals from generalized linear model (GLM), and three simultaneous autoregressive models (SAR_err, SAR_lag, SAR_mix). First row are estimates for species richness and second row for the range size rarity index.*

vascular plant occurrences in the wettest areas and therefore, fewer plants were predicted to occupy these areas with our modeling approach.

The variance of elevation (i.e., elev_var) is the most important factor determining geographical patterns of range size rarity of vascular plants, amphibians, and bats in West Africa (Table 4.2). Although most of the terrain in the study area is flat and homogeneous, those particular areas characterized by an heterogeneous terrain (i.e., mountaineous areas) foster the coexistence of many species with specialized habitats. Temperature related factors are also important determinants of range size rarity of vascular plants and bats, having a positive influence on both, while precipitation has a positive effect only for range size rarity of amphibians.

In general, there is a positive and statistically significant correlation between all possible pair-wise combinations of species richness and the range size rarity of vascular plants, amphibians and bats ($\rho = 0.41 - 0.95$, $P =< 0.001$) (Figure 4.5). That indicates that areas rich in species of any group are areas both, rich in species and rich in endemic species of any other group. Pair-wise correlations between species richness are always higher than pair-wise correlations of the range size rarity between the three groups. The range size rarity of amphibians has the lowest correlations with all other diversity measures, indicating a spatially disjoint distribution of range restricted amphibian species.

Looking in more detail at the scatterplots presented in figure 4.5, discrepancies to the positive and significant correlation mentioned above are found. That is, sites rich in species of one group are not necessarily rich in species of other groups. In fact, general measures of congruence might hide different congruence patterns that are happening at

Table 4.2 – *Environmental determinants of species richness and the range size rarity patterns for vascular plants, amphibians and bats. Estimates presented are the result of the spatial autoregressive model type mixed. Independent and dependent variables were log-transformed. Significance coding: * $p \leq 0.05$;** $p \leq 0.01$;*** $p \leq 0.001$; ns non significant*

	Plants		Amphibians		Bats	
variable	Coefficient	z	Coefficient	z	Coefficient	z
Species Richness						
prec_min	-0.085	-1.04 ns	-0.007	-0.108 ns	-0.091	-1.08 ns
prec	-0.797	-3.05 **	0.491	2.39 *	0.371	1.39 ns
elev_var	0.035	2.64 **	0.128	12.31 ***	0.161	11.98 ***
tmax_min	0.515	6.61 ***	0.216	3.52 ***	0.169	2.13 *
tmin_max	-0.086	-1.15 ns	-0.025	-0.43 ns	0.026	0.34 ns
Range Size Rarity						
prec_min	-0.029	-1.23 ns	0.051	3.61 ***	-0.004	-1.16 ns
prec	-0.353	-4.62 ***	0.049	1.11 ns	-0.007	-0.68 ns
elev_var	0.028	7.39 ***	0.024	10.58 ***	0.007	12.81 ***
tmax_min	0.105	4.61 ***	-0.007	-0.53 ns	0.008	2.69 **
tmin_max	-0.022	-1.00 ns	0.003	0.24 ns	-0.012	-3.95 ***

local scales and smaller extents (Gaston 1996). To illustrate small extent congruence patterns we calculated the correlation between grid cells selected by a moving window of $\approx$ 250 km size. Results show that in fact there is a great amount of local spatial variation between pair-wise congruence of the different groups and diversity measures (Figures 4.6 to 4.8).

The frequency distribution of grid cells with high correlation is higher in all cases (Figures 4.6 to 4.8). However, there are specific areas of negative and no correlation. For example, areas near the coast in Ghana and Ivory Coast have negative correlation coefficients when comparing vascular plants and bats in terms of species richness (figure 4.6) and the range size rarity (figure 4.7). While in those areas there is an increment on the number of vascular plant, there is a decrement on the number of bats. In general, areas rich in species are also areas rich in range-restricted species within any of the taxa studied (Figure 4.8). The only exception to this statement is for bats in some areas in Ivory Coast where correlations are zero or near zero.

Range size rarity hotspots are located at the southern parts of the study area (Figure 4.9). Still, the degree of overlap is rather small (Figure 4.10). The greatest overlap was between amphibians and bats (35.8%). Lower percentage of overlap was obtained between these two taxa and vascular plants (18.9% and 14.2% respectively). Hotspots overlap between all three taxa was 25.5%.

The most important areas (i.e., coincidence of hostpots for the three groups) are found in Nigeria in the limits with Cameroon (i.e., Mont Cameroon), in south-central Ghana, in Ivory Coast in the areas of the Taï National Park and Mont Nimba and in the limits between Liberia and Sierra Leone (Figure 4.10). Hotspots unique for vascular plants can be observed in south-central Nigeria and near the coast in Ivory Coast. Unique amphibian

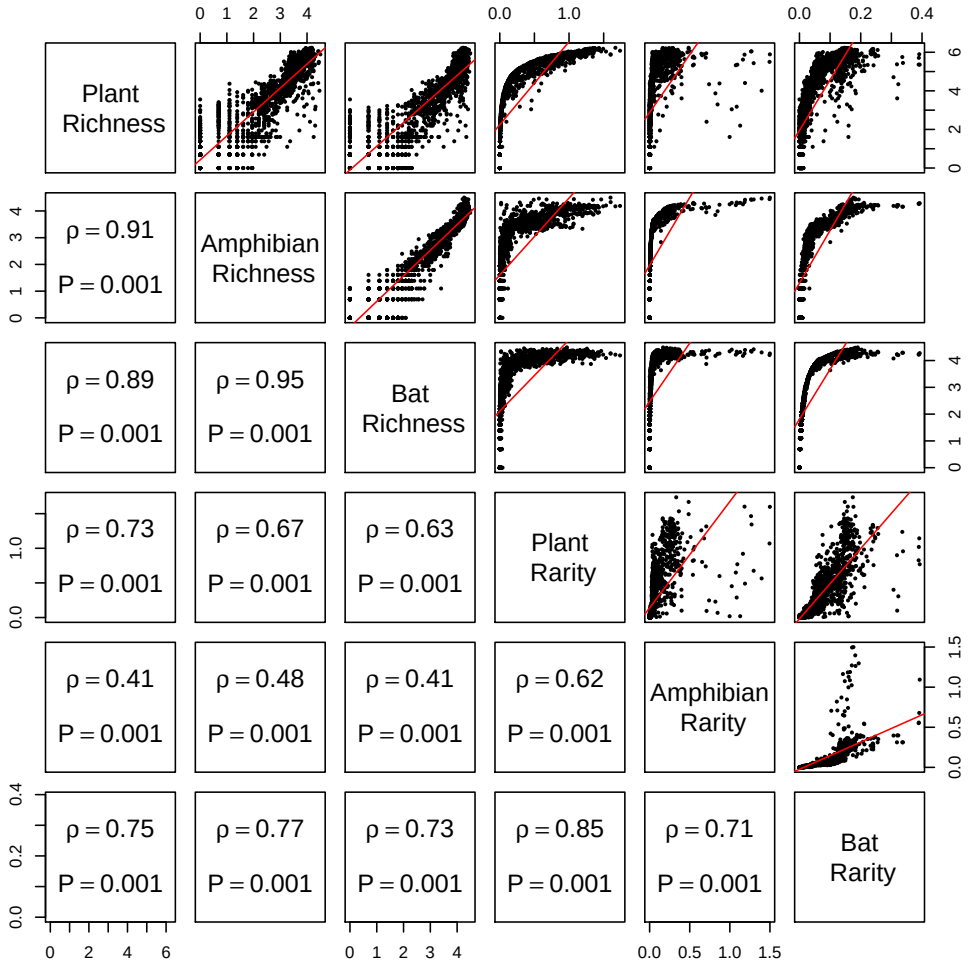

Figure 4.5 – *Pair-wise correlation analysis of species richness and the range size rarity of vascular plants, amphibians and bats. Plots above the diagonal are the correlation scatterplots. The red line shows the best fit of a simple linear regression. Below the diagonal are the rho Spearman correlation values and the significance levels of the null hypothesis of the correlation value being zero.*

hotspots are prominent in Nigeria and for bats in the northernmost areas surrounding the aggregate area of hotspots (Figure 4.10).

In general, less than 50% of the aggregate, common and individual area of vascular plants, amphibians and bats hotspots remain without low degrees of human intervention. Nearly 70% of the most relevant diversity areas (i.e., areas where all taxa hostpots overlap) is affected by some degree of intervention and only 2% is covered by the network of protected areas. Overall, the coverage of protected areas in the hotspots identified in this study is very poor (4.1% to 6.1%). Amphibian hotspots however, are both better covered by protected areas and contain the highest percentage of natural areas (43.6%) (Table 4.3).

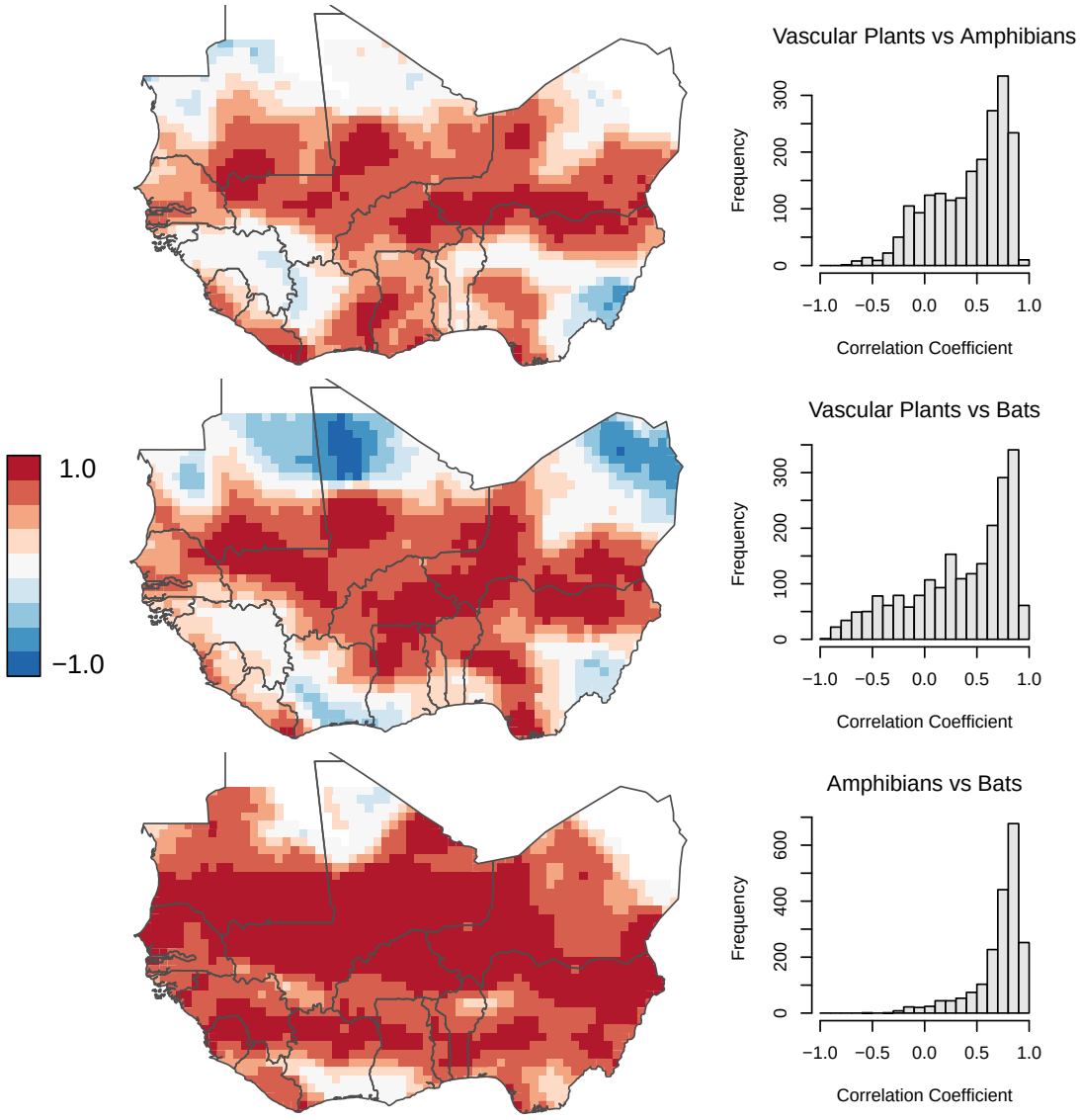

Figure 4.6 – *Small extent variation of the correlation between species richness of vascular plants, amphibians and bats. Correlations were calculated within a moving window of ≈ 250 km radius. Shown to the right of each map are frequency distributions of correlation coefficients.*

4.5 Discussion

We have estimated distribution patterns of species richness and the range size rarity of vascular plants, amphibians, and bats in West Africa. Patterns of vascular plant richness are in agreement with patterns estimated elsewhere (Barthlott et al. 2007, Kier et al. 2006, Barthlott et al. 2005).

But is our approach to model richness distribution patterns the more appropriate? The commonly used approach is what Algar et al. (2009) labeled as "the empirical diversity theory approach". This approach finds the relationship between aggregate species counts and a set of environmental variables on a grid cell basis. Variable coefficients are then

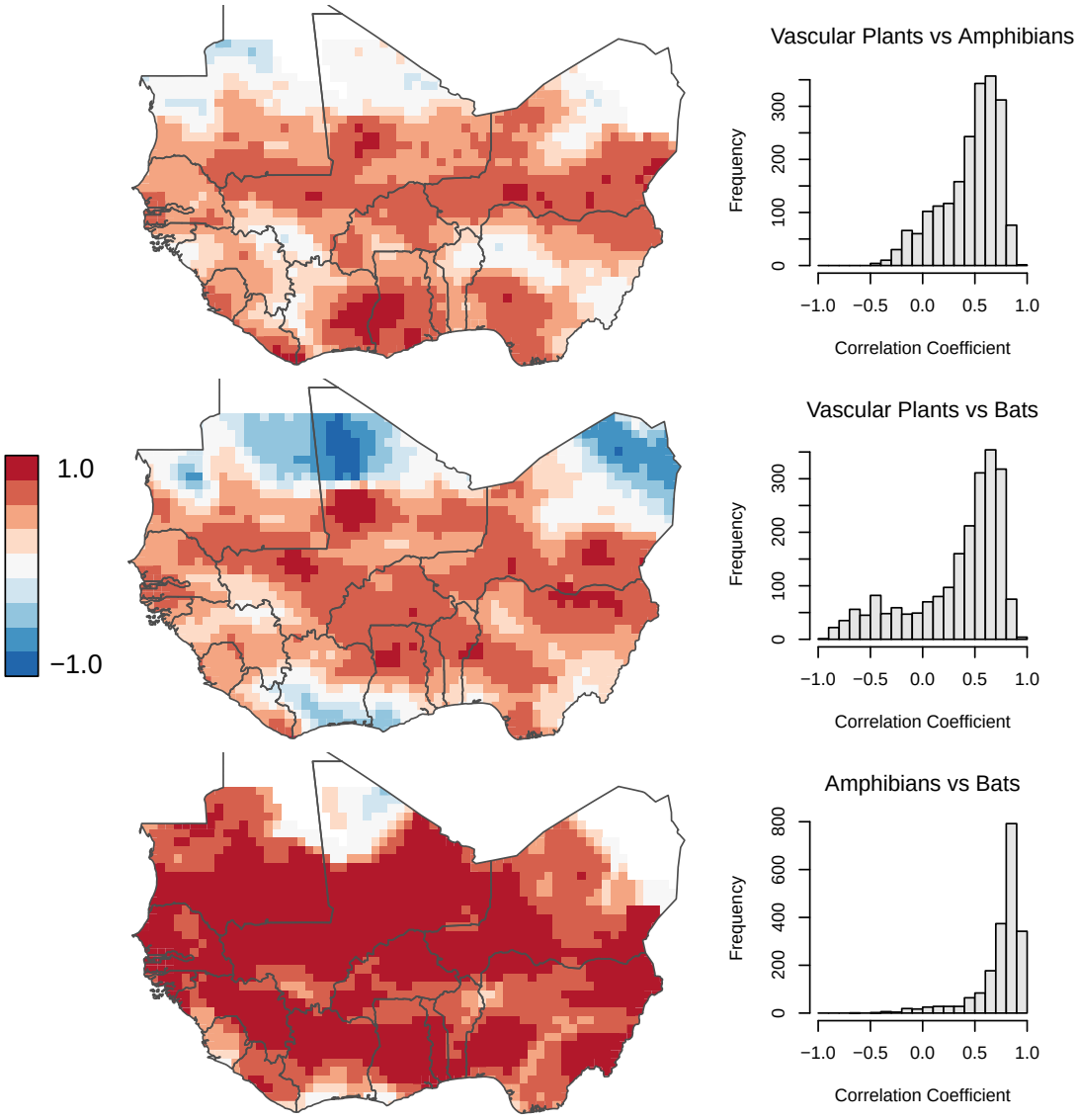

Figure 4.7 – *Small extent variation of the correlation between the range size rarity index of vascular plants, amphibians and bats. Correlation were calculated within a moving window of ≈ 250 km radius. Shown to the right of each map are frequency distributions of correlation coefficients.*

calculated and predictions of species richness can be made (e.g. Kreft and Jetz 2007). In order to produce accurate and reliable results from this approach, a complete census of species is needed. However, aggregate species counts are normally considered as a negative bias estimate of the real number of species in a certain area (see Walther and Moore 2005, their figure 2). Undoubtedly, that is the case in our databases, where aggregate counts of the number of species in a 0.5 degree cell is far from complete. Albeit methodological limitations, we considered species niche modeling as the proper approach to model the incomplete information on species distributions in order to fill up our limited knowledge on species distribution ranges and therefore, to obtain a more realistic picture of richness estimates and spatial patterns. Also, accurate estimates of the range size rarity can only be made with a complete definition of species distribution ranges, which can only be

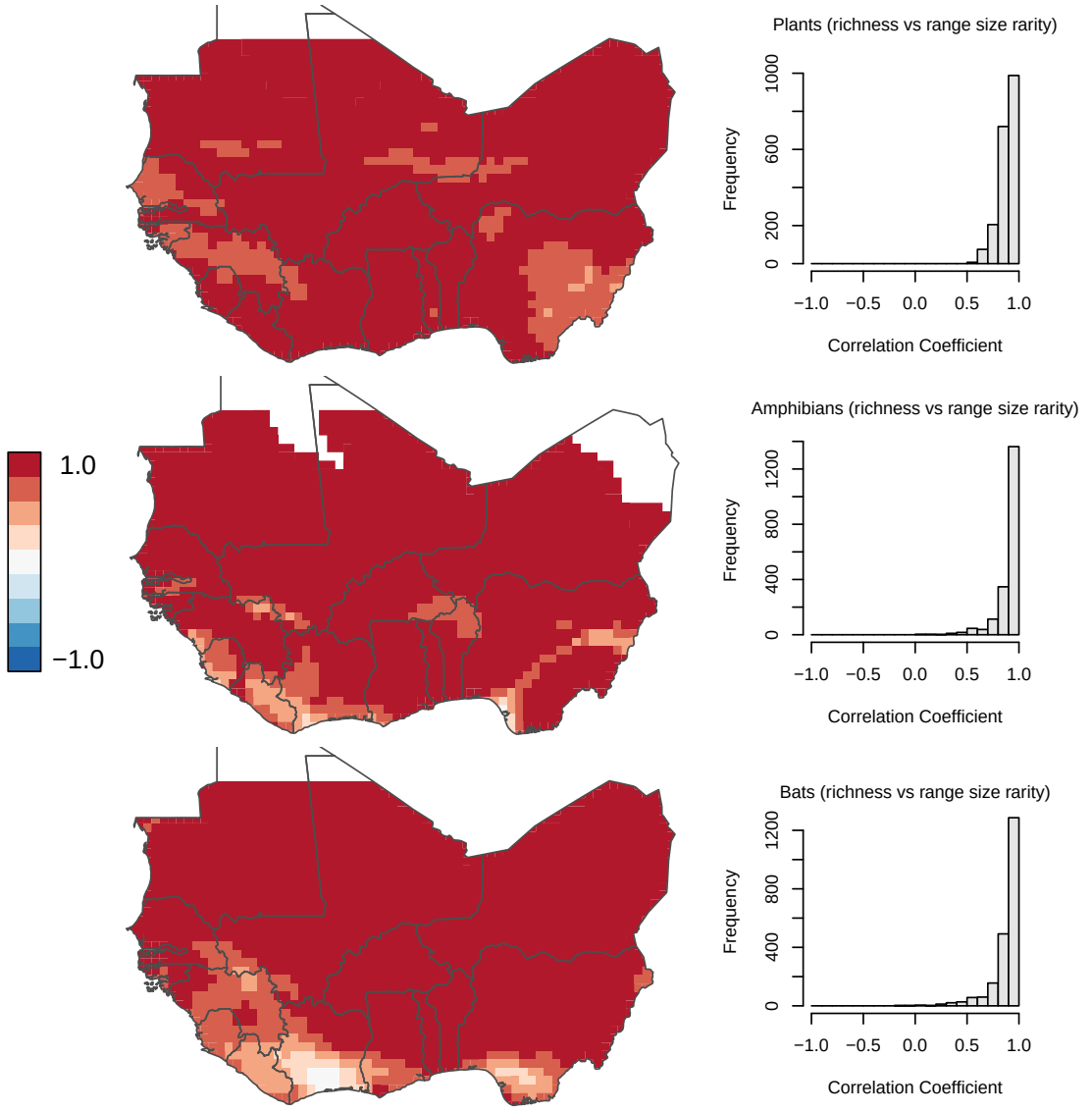

Figure 4.8 – *Small extent variation of the correlation between species richness and the range size rarity of vascular plants, amphibians and bats. Correlations were calculated within a moving window of ≈ 250 km radius. Shown to the right of each map are frequency distributions of correlation coefficients.*

estimated with our approach.

Climate related variables have been constantly considered as main determinants of species richness patterns from regional to global scales (Kreft and Jetz 2007, Hawkins et al. 2003, Currie et al. 2004). Although precipitation and temperature variables were significant determinants of species and range size rarity patterns of vascular plants, amphibians and bats, the variance of elevation values (i.e., elev_var) was the most important variable in all but one case (i.e., for vascular plants) (Table 4.2). The variation in patterns of vascular plants was better described by temperature related factors.

Our results suggest that species richness and range size rarity gradients in West Africa are mainly driven by terrain variability which is in support of the known habitat heterogeneity

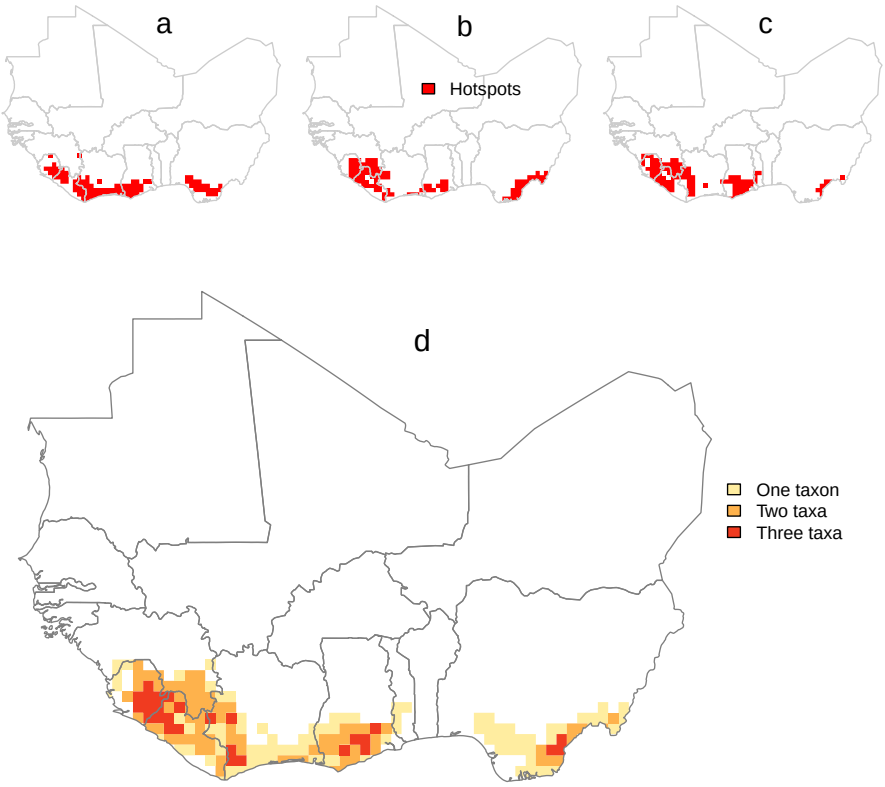

Figure 4.9 – *Maps of selected hotspots for each taxon (a,b and c) and hotspots congruence (d). Hotspots were selected as the number of grid cells with the top 5 percentile of range size rarity values.*

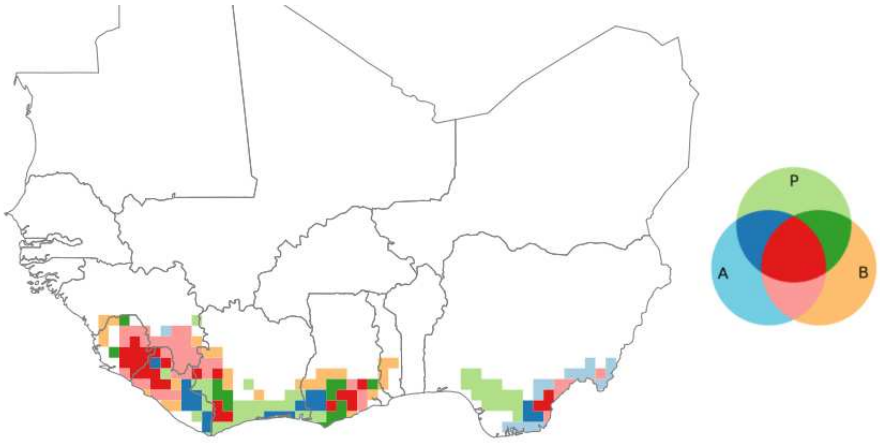

Figure 4.10 – *Geographic distribution of the spatial overlap between hotspot of vascular plants (**P**), amphibians (**A**) and bats (**B**)*

Table 4.3 – *Analysis of land cover (i.e., Natural Areas) and protected areas coverage within the aggregate, common and individual hotspots of vascular plants, amphibians and bats in West Africa. Aggregate hotspots are those areas considered as hotspot at least for one taxa. Common hotspot are areas shared as hotspots for the three taxa.*

	Study Area		Natural Areas		Protected Areas	
Criteria	km	(%)	km	(%)	km	(%)
Aggregate hotspots	583.7	9.3	228.7	39.2	20,8	8.5
Common hotspots	82.5	1.3	25.4	30.8	4.8	2.0
Plant hotspots	324.3	5.0	103.8	32.0	9.9	4.1
Amphibian hotspots	323.8	5.0	141.2	43.6	14.9	6.1
Bats hotspots	323,6	5.0	111.9	34.5	14.2	5.8

hypothesis. This hypothesis proposes that an increase in the number of potential habitats leads to an increase in species diversity in a landscape (MacArthur and MacArthur 1961, Shmida and Wilson 1985, Kerr and Packer 1997, Cramer and Willig 2005). In West Africa the terrain is generally flat, but mountainous areas, where the terrain variability is higher, possibly produce an increment of new available habitats where species turnover increases and more species can exist.

One possible criticism to the way we estimated model coefficients is the recycling of the set of environmental variables first to model each species' distribution ranges and then again to estimate richness and the range size rarity patterns determinants. We believe this approach is valid. Finding richness determinants directly from the databases as they are would be erroneously done since these databases are incomplete and present some degree of spatial bias. In order to calculate the determinants and quantify their influence we need reliable richness patterns. Modeling each species independently makes sense since each species responds differently to environmental conditions By superimposing all individual predictions, a more accurate pattern of richness is obtained. Then, it makes sense to use the same set of environmental factors to see what is defining the richness or range size rarity patterns as a whole. Variable recycling in species distribution modeling applications have been recently applied to improve model performance (Hengl et al. 2009).

Our results suggest that overall pair-wise correlation between patterns of species richness and the range size rarity of vascular plants, amphibians and bats in West Africa is positive and highly significant. Our analysis demonstrates as well that overall pair-wise correlations could lead to over-optimistic conclusions since comparisons at smaller extents varied strongly in intensity. Although areas of positive correlation predominate, there are spots where weak or negative correlations were found. Such divergences in correlation magnitude have also been found in similar studies for other regions (see McKnight et al. 2007). It is then necessary to make a clear delimitation of the spatial extent of analysis if cross-taxon studies are carried out. Only in that way testing ecological hypothesis and taking decisions concerning the conservation of biodiversity can be objectively made.

There is no consensus in the definition of what a hotspot is. In general, a hotspot is defined in relation to the specific diversity aspect under scrutiny or a combination of them. For example, Prendergast et al. (1993) used the top 5% of record-containing 10 km grid cells after ranking all cells by the number of species in each cell (i.e., species richness). McKnight et al. (2007) used only the top 2.5% of 100 km square grid cells containing the highest beta diversity values. We decided to choose the top 5% since the total area obtained (i.e., ≈325 km^2) is similar to the total area of official protected areas in the study area.

Although hotspots identified for each taxa were located near the coast, their congruence was unexpectedly small. Ecological and methodological factors may contribute to this incongruence. Ecologically, each taxon might respond differently to the environmental conditions in the study area.

However, results of species richness and range size rarity determinants indicate that the distribution of three taxa appeared to be explained by similar environmental factors (Table 4.2). Nevertheless, there are slight differences in those factors which we considered as probable causes of distribution incongruence. How congruence patterns will be modify if we include into the analysis other taxa? We expect the overlap between taxa to be lower when we add groups that have notably different ecological requirements. For example, Prendergast et al. (1993) found low congruence between five different biological groups in Britain and attributed these findings to the fact that each group has different ecological requirements (the groups analyzed were butterflies, dragonflies, aquatic plants, breeding birds and liverworts).

Methodologically, factors concerning data collection may be responsible for the low hotspots congruence. Data collection was done independently for each group. Each database was constructed based on different sampling strategies and relying on heterogeneous sources of information. Collection density might have been greater in some particular areas for one the taxon than for the others. We also acknowledged that bias in collection localities (i.e., more collection in accessible areas or in protected areas) might be a second reason for the in-congruence pattern found. For example, we were expecting areas of congruence between all three taxa in the eastern part of Liberia, but since this area has been poorly sampled (i.e. is not accessible) species distribution ranges are underestimated. We forecast that by collecting more data in the existing gaps, species ranges will be more accurate and hotspots overlap will increase.

There are strong differences between the maximum endemism richness values (i.e. range size rarity index) for the three groups. It is higher for plants (4.68) and amphibians (3.48) than for bats (0.48). Bat species in general show wider habitat ranges than vascular plants and amphibians. In the latter two, there is a higher proportion of species with restricted habitat ranges. However, the fact that some vascular plants and amphibians have been sampled in particular areas where collection effort is high may contribute to the high

endemism index, not because the species is restricted but because the sampling collection and information obtained for those species is restricted to a particular area.

If areas for conservation had to be selected based on our results, areas where hotspots of the three taxa overlap will be the best candidate places. Our hotspots are defined based on the range size rarity index which is in turn the result of weighting species richness according to the number of range restricted species on a grid cell basis. However, species richness and range restricted species should not be the only criteria taken into account for conservation purposes, especially when there is a great variation in congruence at local extents. Other diversity aspects that should be considered are the congruence of community similarity or species complementarity (Su et al. 2004, Justus and Sarkar 2002, Bergl et al. 2007) and areas of congruence in beta diversity (i.e. changes in species composition between places)(McKnight et al. 2007).

Human land use is one, if not the main threat to ecosystems and habitats in West Africa (Sanderson et al. 2002, Poorter et al. 2004a). According to our results, the percentage coverage of natural areas within hotspots reached a maximum of 43.6% for amphibians. In general, more than 50% of land within hotspots has some degree of intervention. The situation is much worse for the percentage coverage of protected areas within hotspots, with amphibians hotspots covered at best (i.e., 6.1% coverage). These results corroborate prior findings where the coverage of the network of protected areas in Africa performs poorly in protecting range-restricted species (Burgess et al. 2005). In fact, most of the existing protected areas have been defined based on many criteria but the proper representation of biodiversity (Bergl et al. 2007, Hannah et al. 2007).

Optimizing the prioritization of conservation areas in West Africa might not be that easy after all. Although congruent areas of richness and range-restricted species predominate, there are also areas of weak and negative correlation. In addition, the overlap between hotspots of vascular plants, amphibians and bats is rather small. Being concordance between taxa a main criterion for biodiversity conservation, it remains a challenge for scientist to find the proper methods to deal with incomplete data and fill up information gaps for the efficacy of this criterion as a surrogate for conservation efforts.

CHAPTER 5

GENERAL CONCLUSIONS

Studies and applications in biogeography play a crucial role for the conservation of biodiversity. By understanding how species distribute and the factors that determine their distribution patterns, objective and sound decisions can be made concerning what, where and why to conserve. However, reliable analysis and accuracy of the results are strongly dependent on the quality of the biological databases used, in particular, on the quality and accuracy of the spatial information available. Given the urgent need and existing pressure for strategies for the conservation of biodiversity, scientist and decision makers are obligated to use the existing data from herbarium, museum collections, field relevés despite of its weaknesses (Robertson and Barker 2006). A very first step then must concern a quality assessment of this information (Robertson and Barker 2006, Soberón et al. 2007).

By applying the methodological framework developed in chapter 2 the strengths and weaknesses of biological databases can be quantified. These aspects are visualized in the 'gap selection index' map. The main role of this index is to guide future research into areas that have been poorly investigated in terms of species composition and environmental conditions. Moreover, it serves also the purpose of representing modeling uncertainties when poorly investigated areas are included in biogeographical applications. Maps of species richness or endemism must be displayed together with the site selection index map in order to objectively know where estimates are reliable and where not.

This methodology was applied for the database of vascular plants available for Ivory Coast, Burkina Faso and Benin. Despite of being the most comprehensive database for the region, it was quantitatively demonstrated that there are still large information gaps.

In particular, the distribution of collection effort follows a strongly clustered pattern, which results in very few areas with high collection densities, floristically complete and where the environmental conditions are well represented. Further, the factors that might potentially be responsible for the biased pattern of collection localities were identified. For example, research has been focused near the coast and near to the main roads in Ivory Coast while in Burkina Faso the Sahelian zone has received extraordinaire attention. In addition, this collection distribution bias represents environmental bias as well, that is, some environmental conditions are over-represented and some under-represented in the distribution of collection localities.

By relating some of the factors considered in the evaluation of the database, important conclusions can be drawn. For example, the relation between density of collection localities and floristic inventory completeness on a grid cell basis makes it evident that little collection effort is needed in order to obtain a complete floristic inventory on a 10 km^2 grid cell. Another important advantage of the developed methodology is that it is scale independent. All procedures can be applied at different spatial scales, allowing to identify at which scale the data fulfill the assumptions required for any modeling approach. In general, results of these analysis allow scientist to better focus their efforts to fill up information gaps when performing additional inventories.

The results of using the spatial information of this database as it is, for example to model and predict the ecological niche of species, without any consideration of the spatial bias constraint, could be equally biased. Prediction will more precisely represent collection effort rather than the real habitat ranges of species. Knowing the possible causes of bias in the data and its spatial distribution is crucial to develop and identify proper methodologies and techniques to deal with this constraint.

In fact, the constraint found in this database is very common in such biological databases and several investigations have proposed methodologies to tackle this situation (see Zaniewski et al. 2002, Kadmon et al. 2004, Phillips et al. 2009, for examples). Three different approaches were tested in this thesis (chapter 3), in the context of using ecological niche models to estimate species distribution ranges and species richness patterns. The first approach is the common selection of random background data from the study area (i.e., 'random-background'). The second approach limits the selection of background data to the same sites of collection localities, except those where the target species is documented (i.e., 'target-background', Phillips et al. (2009)). The third approach selects background data based on a weighted function of the 'site selection index' developed in chapter 2 (i.e., index-background). Model performance improved significantly when using 'target-background' and species richness estimates were more accurate to the known richness gradient in the area. Non-significant differences were found in cases where the number of occurrence records reach 100 observation. One of the priorities for future research is further collection of those species that at present are underrepresented in the database.

The database evaluation procedure proposed in this thesis explicitly exposed the deficiencies and flaws of the database used as a study case. At the same time, it gives advice as to what and where the future tasks should be to improve its quality. In spite of its deficiencies, the information contained in the database must be used for biogeographical application that lead to the conservation of biodiversity. There has been a large amount of scientific research that can be used to mobilize the information available, despite of its weaknesses. Using methodologies and modeling strategies, like the ones applied in this thesis, are the best example to demonstrate that still accurate results can be obtained. These results can contribute to straighten the focus of biodiversity conservation actions.

Species distribution models (SDM) are nowadays one of the main tools of analysis in biogeographical applications. With the increase in cost-free data and software, generating predictions of species ranges appear to be a quite easy task. However, we have demonstrated how results of applying SDMs can result in misleading information. If such results are used as basis for decision making regarding the conservation of biodiversity, or the criteria to test ecological and evolutionary hypotheses, the outcomes may contain substantial errors. We recommend a critical use of SDMs by comparing different strategies in order to minimize prediction uncertainties and obtain the best possible results.

Different approaches exist to make use of the information available in biological databases to estimate species richness patterns. Although taxon-based approaches (e.g., ecological niche models) are not suitable for areas with limited data availability (Mutke and Barthlott 2005), it was demonstrated that this approach successfully made use of the information of databases containing distributional data on vascular plants, amphibians, and bats to estimate species distribution ranges and therefore, species richness patterns. The final distribution patterns obtained are in concordance with the known species richness gradients in West Africa (Barthlott et al. 2005, Kreft and Jetz 2007).

In agreement with theories claiming the role of habitat heterogeneity as a main determinant of species richness gradients (Shmida and Wilson 1985, Kerr and Packer 1997, Rahbek and Graves 2001), this thesis has found this criterion as one of the main factors governing the spatial variation of overall species richness and of range restricted species. Areas of high species richness occur in areas of high elevation heterogeneity (the latter can be considered as a surrogate of habitat heterogeneity). These areas correspond mainly to mountainous regions in West Africa, where climatic conditions, as well as soil and geology characteristics change at a steep rate (Braun et al. 2002). Since the environmental conditions in these areas differ from the homogeneous conditions that predominate in the study area, the existence of a higher proportion of range restricted species was expected.

Evaluating overall congruence between species richness patterns of different taxa can result in misleading information. In this thesis high positive overall correlations were found in a pair-wise comparison of both species richness and the range size rarity index between

vascular plants, amphibians and bats. The first conclusion from this result is that by protecting areas rich in vascular plant species, also areas of high amphibians and bats species will be protected. However, correlation analyzes done at local scales resulted in areas also of weak and negative correlations. Accordingly, some areas rich in species of one group are poor in species of other groups.

Conclusions and decisions towards the conservation of overall biodiversity must take into account the local spatial variability of species richness and endemism of any biological group when estimating patterns of congruence between them. In fact, the total overlap of areas rich in species and range restricted species between the three groups analyzed was very small. Prioritization of conservation of those restricted areas is an urgent matter, since the amount of remnant natural vegetation and the coverage of international recognized protected areas in this areas is minimal.

Certainly, the overarching conclusion of this thesis is that in spite of the availability of biological databases and tools for biogeographical analyses and applications, results might be harm by a series of constraints regarding data quality, model assumptions and scale of analysis. Any biogeographical study should follow an objective and scientifically sound analysis of the quality of the information contained in biological databases in order to know where are the deficiencies and how to improve its quality. At the same time, several modeling techniques or approaches known to be robust against the constraints in the data should be used and compared to minimize result uncertainties. Finally, analysis at different spatial scales facilitate the identification of patterns that otherwise will be hidden and that are key components for the decisions concerning the conservation of biodiversity.

CHAPTER 6

SUMMARY

García Márquez, Jaime Ricardo (2010). ***Biogeographical Analyses and Applications: The Study of Vascular Plant Distribution Patterns in West Africa.*** Doctoral thesis, Mathematisch–Naturwissenschaftliche Fakultät, Rheinische Friedrich–Wilhelms-Universität Bonn.

Studies and applications in biogeography aim at understanding the causes and determinants of past, current, and future diversity distribution patterns. In this context, niche models are currently important tools. They make use of species georeferenced locations and sets of environmental variables to determine the potential suitable habitats of species. The derived results are applied for the conservation of biodiversity but their reliability strongly relies on the quality of the spatial information contained in biological databases. Biodiversity conservation effectiveness strongly depends on the identification of congruent areas for a set of diversity indicators. For example, congruent areas of high species richness for different biological groups.

In the first part of this thesis a methodological framework is developed to evaluate the quality of biological databases. As a case example, a database of vascular plants in Benin, Ivory Coast and Burkina Faso is analyzed. It consists of a total of 4,587 species collected in 2,931 different localities.

Main criteria to evaluate database quality are the spatial configuration of collection localities, their spatial and environmental bias, and their floristic inventory completeness. It was shown that collection localities are unevenly distributed, forming strong clustered patterns and concentrated in the vicinity to cities, the coast, rivers, roads and protected

areas. Collection localities represent only a narrow range of the environmental conditions in the study area. A gap selection index was created integrating density of collection localities, environmental bias and inventory completeness to represent those areas in need of more information and to guide future research activities. According to the gap selection index, only the Sahelian zone and surroundings of Comin-Yanga city in Burkina Faso, areas close to the coast and in the border with Liberia in Ivory Coast, and the region of the eastern Guinean forest in Benin are good represented.

Different niche models strategies are tested to compare model performance between commonly used approaches, and those that seek to minimize the influence of spatial bias present in the occurrence records. The maximum entropy technique (i.e., Maxent) is used to model species ranges based on biased occurrence records and three types of environmental background information sets. These are random locations ('random background'), the locations of all collection localities except those of the target species ('target background'), and random locations weighed as a function of the gap selection index ('index background'). In average, target background based models perform better than random and index background based models. A visual examination confirms that only target background based models are able to correct for spatial bias. Species richness patterns are estimated to qualitatively describe spatial variations as a result of the three approaches employed Richness patterns strongly vary across the three approaches but those based on target background models are more accurate to the known positive north-south richness gradient in the study area.

In the second part of this thesis, the most comprehensive databases of vascular plants, amphibians and bats in West Africa are used to estimate geographical patterns of species richness and the range size rarity index (as a surrogate of species richness and endemism) at a half degree resolution. The relationship between abiotic factors and these two aspects of diversity are examined through the use of spatial auto-regressive techniques, indicating that elevation heterogeneity and temperature are the main determinants of variation in species richness and range size rarity patterns of the three taxa. Pair-wise correlation comparisons between the three groups for species richness and the range size rarity showed generally high and significant correlations (i.e., > 0.9). However, weak and even negative correlations were found when the comparison was applied at small geographical extents. Hotspots of the range size rarity index for the three groups together occupy 9.3% (i.e. 583.7 km^2) of the study area. 39.2% of this area is still covered by natural vegetation and 8.5% is covered by protected areas.

The main finding of this thesis is that despite the large amount of information available to carry out research in biogeography, analysis and application should not be done without an objective and sound analysis of data quality. Only in this way sources of error and uncertainty can be identified, and proper modeling techniques can be applied. These are crucial steps that need to be followed before drawing conclusions and making decisions for the conservation of biodiversity.

CHAPTER 7

ZUSAMMENFASSUNG

García Márquez, Jaime Ricardo (2010). ***Biogeographische Analyse und ihre Anwendung: Eine Studie der Verbreitungsmuster von Gefä"spflanzen in Westafrika.*** Dissertation, Mathematisch–Naturwissenschaftliche Fakult"at, Rheinische Friedrich–Wilhelms-Universität Bonn.

Biogeographische Studien und sich daraus ableitende Anwendungen haben zum Ziel, die Ursachen und Faktoren historischer, aktueller und zukünftiger Verbreitungsmuster zu verstehen. In diesem Zusammenhang finden ökologische Nischenmodelle derzeit vielfältige Anwendung. Diese nutzen georeferenzierte Standorte und eine Reihe vom Umweltvariablen, um zu analysieren, welche Lebensräume für bestimmte Arten potenziell geeignet sind. Die gewonnenen Ergebnisse lassen sich im Naturschutz für den Erhalt von Biodiversität nutzen. Ihre Glaubwürdigkeit beruht jedoch stark auf der Qualität der räumlichen Informationen, die in biologischen Datenbänken enthalten sind. Die Effizienz von Ma"snahmen zur Biodiversitätserhaltung hängt allerdings stark davon ab, ob in den priorisierten Gebieten mehrere Diversitätsindikatoren kombiniert auftreten, beispielsweise ob Zentren der Artenvielfalt für unterschiedliche taxonomische Gruppen zusammenfallen.

Im ersten Teil dieser Arbeit wird ein methodisches Ger"ust erstellt, um die Qualit"at von biologischen Datenb"anken zu bewerten. Als Fallbeispiel wird eine Datenbank zur Verbreitung von Gef"a"spflanzen in Benin, der Elfenbeink"uste und in Burkina Faso analysiert. Sie enth"alt insgesamt 4.587 Arten, die an 2.931 verschiedenen Orten gesammelt wurden.

Hauptkriterien f"ur die Bemessung der Qualit"at von Datenbanken sind die r"aumliche Lage der Fundpunkte sowie deren geographische und taxonomische Unvollst"andigkeit.

Es hat sich gezeigt, dass die Fundpunkte in der Tat ungleichm”a”sig verteilt sind, und sich vermehrt in der Umgebung von St”adten, der K”uste, Fl”ussen, Stra”sen und Schutzgebieten befinden. Die Fundpunkte repr”asentieren au”serdem nur einen Teilbereich der im Untersuchungsgebiet vorkommenden Umweltbedingungen. Es wurde ein Index erstell, der die Dichte der Fundpunkte, die abgedeckten Umweltbedingungen und die taxonomische Vollst”andigkeit der Daten kombiniert (Gap Selection Index), um diejenigen Gebiete aufzuzeigen, die nach diesen Kriterien unterbesammelt sind und eine bessere Inventarisierung erfordern. Gem”a”s dieses Indexes, sind nur die Sahelzone und die Umgebung der Stadt Comin-Yanga in Burkina Faso, Gebiete in K”ustennähe im Grenzgebiet zu Liberia in der Elfenbeink”uste und die Region der ”ostlichen guineischen W”alder in Benin gut in der Datenbank repr”asentiert.

Es werden verschiedene Nischenmodellierungsverfahren genutzt, um die Ergebnisse etablierter und neu entwickelter Ans”atze zu vergleichen, die die r”aumliche Unausgewogenheit von Datens”atzen mit einbeziehen. Die Maximum-Entropie-Technik (z.B. Maxent) wurde genutzt, um Artverbreitungen mit unvollst”andigen Fundpunktdaten nach drei verschiedenen Ans”atzen zu modellieren. Die Ans”atze unterscheiden sich nach der Methode, mit der die f”ur die Begrenzung der Areale n”otige Pseudoabsenzen generiert werden. Diese drei Ans”atze sind eine Zufallsauswahl ”‘random background”’, einen Konzentration der Pseudoabsenzen in m”oglichst gut besammelten Gebieten, in denen die Art aber nicht dokumentiert ist ”‘target background”’, und Absenzen, deren Auswahl mit dem o.g. ”‘Gap Selection Index”’ gewichtet wurde, ”‘index background”’. Im Mittel haben hierbei die Modelle mit ”‘target background”’ am besten abgeschnitten. Eine visuelle ”Uberpr”ufung best”atigt, dass nur diese Modelle angemessen die r”aumlichen Datenl”ucken ausgleichen. Auf Basis der verschiedenen Ans”atze wurden Artenvielfaltmuster berechnet und die Unterschiede der Ergebnisse qualitativ beschrieben. Die Artenvielfaltmuster variieren stark zwischen den drei Ans”atzen. Diejenigen Modelle, die auf ”‘target background”’ basieren, bilden den bekannten positiven Nord-S”ud-Gradienten im Forschungsgebiet jedoch am besten ab.

Im zweiten Teil dieser Arbeit wurden die umfangreichsten Datenbanken zur Verbreitung von Gef”a”spflanzen, Amphibien und Flederm”ausen in Westafrika genutzt, um r”aumliche Muster der Artenvielfalt und den Endemismusreichtum (als kombinierten Index von Artenvielfalt und Endemismus) auf einer 0,5-Grad-Aufl”osung zu berechnen. Mit Hilfe von r”aumlichen auto-regressiven Verfahren wurde die Beziehung zwischen diesen Diversit”atsma”sen und verschiedenen abiotischen Faktoren untersucht, die aufzeigt, dass die Topographie und die Temperatur die wichtigsten erkl”arenden Variablen bez”uglich der Unterschiede der Artenvielfalt und des Endemismusreichtums dieser drei Taxa darstellen. Paarweise Korrelationen zwischen den drei Gruppen f”ur Artenvielfalt und Endemismusreichtum haben grunds”atzlich starke und signifikante Zusammenh”ange gezeigt ($r > 0{,}9$). F”ur den Vergleich der Muster auf lokaler Ebene waren diese Zusammenh”ange allerdings deutlich geringer oder sogar negativ. Zentren des Endemismusreichtums f”ur die drei

Gruppen nehmen 9,3% (d.h. 583,7 km^2) des Forschungsgebietes ein. 39,2% dieses Gebietes ist von nat"urlicher oder naturnaher Vegetation bedeckt und 8,5% sind Schutzgebiete.

Die wichtigste Erkenntnis dieser Arbeit ist, dass trotz der gro"sen Menge an vorhandenen Daten im Kontext von biogeographischen Analysen eine klare Zielsetzung und eine gr"undliche Untersuchung der Datenqualit"at unabdingbar ist. Nur auf diese Weise k"onnen m"ogliche Fehlerquellen und Unsicherheiten festgestellt und passende Modelltechniken angewandt werden. Dies ist eine Grundvoraussetzung f"ur die Verwendung von so gewonnenen Modellergebnissen in der Naturschutzplanung.

BIBLIOGRAPHY

Ake Assi, L., 2001. Flore de la Côte d'Ivoire: catalogue systematique, biogeographie et ecologie I. Biossiera 57. Editions de Conservatoire et Jardin Botanique, Geneva.

Ake Assi, L., 2002. Flore de la Côte d'Ivoire: catalogue systematique, biogeographie et ecologie II. Biossiera 58. Editions de Conservatoire et Jardin Botanique, Geneva.

Algar, A. C., Kharouba, H. M., Young, E. R., Kerr, J. T., 2009. Predicting the future of species diversity: macroecological theory, climate change, and direct tests of alternative forecasting methods. Ecography 32, 22–33.

Allouche, O., Steinitz, O., Rotem, D., Rosenfeld, A., Kadmon, R., 2008. Incorporating distance constraints into species distribution models. Journal of Applied Ecology 45, 599–609.

Anderson, R. P., Lew, D., Peterson., A. T., 2003. Evaluating predictive models of species' distributions: criteria for selecting optimal models. Ecological Modelling 162, 211–232.

Araújo, M. B., Guisan, A., 2006. Five (or so) challenges for species distribution modelling. Journal of Biogeography 10, 1677 – 1688.

Araújo, M. B., New, M., 2006. Ensemble forecasting of species distributions. Trends in Ecology and Evolution 22, 42–47.

Araújo, M. B., Whittaker, R. J., Ladle, R. J., Erhard, M., 2005. Reducing uncertainty in projections of extinction risk from climate change. Global Ecology and Biogeography 14, 529–538.

Austin, M., 2002. Spatial prediction of species distribution: An interface between ecological theory and statistical modelling. Ecological Modelling 157, 101–118.

Baddeley, A., Möller, J., Waagepetersen, R., 2000. Non- and semi-parametric estimation of interaction in inhomogeneous point patterns. Statistica Neerlandica 54, 329–350.

Baddeley, A., Turner, R., 2005. Spatstat: an R package for analyzing spatial point patterns. Journal of Statistical Software 12 (6), 1–42, ISSN 1548-7660.
URL `www.jstatsoft.org`

Balmford, A., Crane, P., Dobson, A., Green, R. E., Mace, G. M., 2005. The 2010 challenge: data availability, information needs and extraterrestrial insights. Philosophical Transactions of the Royal Society 360, 221–228.

Barthlott, W., Biedinger, N., Braun, G., Feig, F., Kier, G., Mutke, J., 1999. Terminological and methodological aspects of the mapping and analysis of global biodiversity. Acta Botanica Fennica 162, 103–110.

Barthlott, W., Hostert, A., Kier, G., Küper, W., Kreft, H., Mutke, J., Rafiqpoor, M. D., Sommer, J. H., 2007. Geographic patterns of vascular plant diversity at continental to global scales. Erdkunde 61, 305–315.

Barthlott, W., Lauer, W., Placke, A., 1996. Global distribution of species diversity in vascular plants: towards a world map of phytodiversity. Erdkunde 50, 317–328.

Barthlott, W., Mutke, J., Rafiqpoor, M. D., Kier, G., Kreft, H., 2005. Global centres of vascular plant diversity. Nova Acta Leopoldina 92, 61–83.

Barthlott, W., Mutke, J., Sommer, J. H., Küper, W., 2003. Biodiversity of africa in the global context: Spatial patterns of vascular plant diversity in a changing environment. In: Sustainable use and conservation of biological diversity - A challenge for society. Proceedings of the International Symposium Berlin, 1st - 4th December 2003. Bonn, pp. 166–167.

Bartholomé, E., Belward, A. S., 2005. Glc2000: A new approach to global land cover mapping from earth observation data. International Journal of Remote Sensing 26, 1959–1977.

Baselga, A., Novoa, F., 2006. Diversity of chrysomelidae (coleoptera) in galicia, northwest spain: Estimating the completeness of the regional inventory. Biodiversity and Conservation 15, 205–230.

Berghaus, H., 1849. Physikalischer Atlas. Perthes Justus (Gotha).

Bergl, R. A., Oates, J. F., Fotso, R., 2007. Distribution and protected area coverage of endemic taxa in west africa's biafran forests and highlands. Biological Conservation 134 (2), 195 – 208, conservation in Areas of High Population Density in Sub-Saharan Africa.

Berman, M., Diggle, J., 1989. Estimating weighted integrals of the second-order intensity of a spatial point process. Journal of the Royal Statistical Society 51, 81–92.

Bestelmeyer, B., Miller, J., Wiens, J., 2003. Applying species diversity theory to land management. Ecological Applications 13, 1750–1761.

Bivand, R., with contributions by Luc Anselin, ao, R. A., Berke, O., Bernat, A., Carvalho, M., Chun, Y., Christensen, B., Dormann, C., Dray, S., Halbersma, R., Krainski, E., Lewin-Koh, N., Li, H., Ma, J., Millo, G., Mueller, W., Ono, H., Peres-Neto, P., Reder, M., Tiefelsdorf, M., , Yu., D., 2009. spdep: Spatial dependence: weighting schemes, statistics and models. R package version 0.4-36.
URL `http://CRAN.R-project.org/package=spdep`

Bonn, A., Schröder, B., 2001. Habitat models and their transfer for single and multi-species groups: a case study of carabids in an alluvial forest. Ecography 24, 483 – 496.

Boyce, M. S., 1992. Population viability analysis. Annual Review of Ecology and Systematics 23, 481– 506.

Braun, G., Mutke, J., Reder, A., Barthlott, W., 2002. Biotope patterns, phytodiversity and forestline in the andes, based on gis and remote sensing data. In: Mountain Biodiversity: a global assessment. London, pp. 75–89.

Brooks, T. M., Balmford, A., Burgess, N. D., Fjeldså, J., Hansen, L. A., Moore, J., Rahbek, C., Williams, P. H., August 2001. Toward a blueprint for conservation in africa. Bioscience 51, 613–624.

Brown, J. H., Lomolino, M. V., 1998. Biogeography, 2nd Edition. Sinauer Associates, Sunderland.

Brown, J. H., Maurer, B. A., 1989. Macroecology: The division of food and space among species on continents. Science 243, 1145–1150.

Burgess, N. D., Küper, W., Mutke, J., Brown, J., Westaway, S., Turpie, S., Meshack, C., Taplin, J. R. D., McClean, C., Lovett, J. C., 2005. Major gaps in the distribution of protected areas for threatened and narrow range afrotropical plants. Biodiversity and Conservation 14, 1877–1894.

Burnham, K. P., Overton, W. S., 1979. Robust estimation of population size when capture probabilities vary among animals. Ecology 60, 927–936.

Chapin, F. S., Zavaleta, E. S., Eviner, V. T., Naylor, R. L., Vitousek, P. M., Reynolds, H. L., Hooper, D. U., Lavorel, S., Sala, O. E., Hobbie, S. E., Mack, M. C., Diaz, S., 2000. Consequences of changing biodiversity. Nature 405, 234–242.

Chatelain, C., Dao, H., Gautier, L., Spichiger, R., 2004. Biodiversity of West African Forests. An Ecological Atlas of Woody Plant Species. CABI, Ch. Forest cover changes in Côte d'Ivoire and Upper Guinea, pp. 15–32.

Chatelain, C., Gautier, L., Spichiger, R., 2001. Application du sig ivoire à la distribution potentielle des espèces en fonction des facteurs écologiques. Systematics and Geography of Plants 71, 313–326.

Chesson, P., 2000. Mechanisms of maintenance of species diversity. Annual Review of Ecology and Systematics 31, 343–366.

Colwell, R. K., Coddington, J. A., 1994. Estimating terrestrial biodiversity through extrapolation. Phil. Trans. R. Soc. Lond 345, 101–118.

Conroy, G. C., Anemone, R. L., Regenmorter, J. V., Addison, A., 2008. Google earth, gis, and the great divide: A new and simple method for sharing paleontological data. Journal of Human Evolution 55 (4), 751 – 755.

Cramer, M. J., Willig, M. R., 2005. Habitat heterogeneity, species diversity and null models. Oikos 108, 209–218.

Cressie, N. A., 1993. Statistics for spatial data. Wiley Series in Probability and Mathematical Statistic.

Currie, D. J., Mittelbach, G., Cornell, H. V., Field, R., Guégan, J., Hawkins, B. A., Kaufman, D. M., Kerr, J. T., Oberdorff, T., O'Brien, E. M., Turner, J. R. G., 2004. Predictions and tests of climate-based hypotheses of broad-scale variations in taxonomic richness. Ecology Letters 7, 1121–1134.

Currie, D. J., Paquin, V., 1987. Large-scale biogeographical patterns of species richness of trees. Nature 329, 326–327.

de Leeuw, J., Jia, H., Yang, L., Liu, X., Schmidt, K., Skidmore., A. K., 2006. Comparing accuracy assessments to infer superiority of image classification methods. International Journal of Remote Sensing 27, 223–232.

De Marco Jr, P., Felizola Diniz-Filho, J. A., Bini, L. M., 2008. Spatial analysis improves species distribution modelling during range expansion. Biology letters 0210, 1–4.

Defense Mapping Agency (DMA), 1992. Digital chart of the world. Tech. rep., Defense Mapping Agency, Fairfax, Virginia.

Díaz, S., Fargione, J., Chapin III, F. S., Tilman, D., 2006. Biodiversity loss threatens human well-being. PLoS Biology 4, 1300–1305.

Diggle, P., 2003. Statistical Analysis of Spatial Point Patterns, 2nd Edition. Arnold Publishers.

Dirzo, R., Raven, P. H., 2003. Global state of biodiversity and loss. Ann. Rev. Environ. Resour: 28, 137–167.

Dormann, C., Purschke, O., García Márquez, J. R., Lautenbach, S., Schröder, B., 2008. Components of uncertainty in species distribution analysis: a case study of the great grey shrike. Ecology 89, 3371–3386.

Dormann, C. F., 2007a. Effects of incorporating spatial autocorrelation into the analysis of species distribution data. Global Ecology and Biogeography 16, 129–138.

Dormann, C. F., 2007b. Promising the future? global change projections of species distributions. Basic and Applied Ecology 8, 387–397.

Dormann, C. F., McPherson, J. M., Araújo, M. B., Bivand, R., Bolliger, J., Carl, G., Davies, R. G., Hirzel, A., Jetz, W., Kissling, W. D., Kühn, I., Ohlemüller, R., Peres-Neto, P. R., Reineking, B., Schröder, B., Schurr, F. M., Wilson, R., 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography 30, 609–628.

Edwards, J. L., Lane, M. A., Nielsen, E. S., 2000. Interoperability of biodiversity databases: Biodiversity information on every desktop. Science 289, 2312.

Elith, J., Burgman, M. A., Regan, H. M., 2002. Mapping epistemic uncertainties and vague concepts in predictions of species distribution. Ecological Modelling 157, 313–329.

Elith, J., Graham, C. H., Anderson, R. P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R. J., Huettmann, F., Leathwick, J. R., Lehmann, A., Li, J., Lohmann, L. G., Loiselle, B. A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J. M. M., Peterson, A. T., Phillips, S. J., Richardson, K., Scachetti-Pereira, R., Schapire, R. E., Soberón, J., Williams, S., Wisz, M. S., Zimmermann, N. E., 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29, 129–151.

Engler, R., Guisan, A., Rechsteiner, L., 2004. An improved approach for predicting the distribution of rare and endangered species from occurence and pseudo-absence data. Journal of Applied Ecology 41, 263–274.

Faith, D., Walker, P., 1996. Environmental diversity: on the best-possible use of surrogate data for assessing the relative biodiversity of sets of areas. Biodiversity and Conservation 5, 399–415.

Farber, O., Kadmon, R., 2003. Assessment of alternative approaches for bioclimatic modeling with special emphasis on the mahalanobis distance. Ecological Modelling 160, 115–130.

Fielding, A. H., Bell, J. F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation 24, 38–49.

Fisher, W. D., 1958. On grouping for maximum homogeneity. Journal of the American Statistical Association 53, 789–798.

Fleishman, E., Noss, R. F., Noon, B. R., 2006. Utility and limitations of species richness metrics for conservation planning. Ecological Indicators 6 (3), 543 – 553.

Fortin, M.-J., Dale, M., 2005. Spatial Analyses: A Guide for Ecologist. Cambridge University Press.

Franklin, J., 1995. Predictive vegetation mapping: geographic modelling of biospatial patterns in relation to environmental gradients. Progress in Physical Geography 19, 474–499.

Freitag, S., Hobson, C., Biggs, H. C., Jaarsveld, A. S. V., 1998. Testing for potential survey bias: the effect of roads, urban areas and nature reserves on a southern african mammal data set. Animal Conservation 1, 119–127.

Friis, I., 1999. Mapping the african flora – trends in the development of methods and applications. In: African Plants: Biodiversity, Taxonomy and Uses. Kew, pp. 131–151.

Funk, V., Richardson, K., 2002. Systematic data in biodiversity studies: use it or lose it. Systematic Biology 51, 303–16.

Funk, V., Richardson, K. S., Ferrier, S., 2005. Survey-gap analysis in expeditionary research: where do we go from here. Biological Journal of the Linnean Society 85, 549 – 567.

Funk, V., Zermoglio, M., Nasir, N., 1999. Testing the use of specimen collection data and gis in biodiversity exploration and conservation decision making in guyana. Biodiversity and Conservation 8, 727–751.

Gaston, K. J., 1996. Biodiversity - congruence. Progress in Physical Geography 20, 105–112.

Gautier, L., Aké Assi, L., Chatelain, C., Spichiger, R., 1999. Sig ivoire: A geographic information system for biodiversity management in ivory coast. In: African Plants: Biodiversity, Taxonomy and Uses. Kew, pp. 183–194.

Gotelli, N. J., Colwell, R. K., 2001. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. Ecology Letters 4, 379–391.

Graham, C., Moritz, C., Williams, S. E., 2006. Habitat history improves prediction of biodiversity in rainforest fauna. PNAS 103, 632 – 636.

Graham, C. H., Ferrier, S., Huettman, F., Moritz, C., Peterson, A. T., 2004. New developments in museum-based informatics and applications in biodiversity analysis. Trends in Ecology and Evolution 19, 497–503.

Graham, M. H., 2003. Confronting multicollinearity in ecological multiple regression. Ecology 84 (11), 2809–2815.

GRASS Development Team, 2008. Geographic Resources Analysis Support System (GRASS GIS) Software. Open Source Geospatial Foundation.
URL `http://grass.osgeo.org`

Guisan, A., Broennimann, O., Engler, R., Vust, M., Yoccoz, N., Lehman, A., Zimmermann., N., 2006. Using niche-based models to improve the sampling of rare species. Conservation Biology 20, 501 – 511.

Guisan, A., Zimmermann, N. E., 2000. Predictive habitat distribution models in ecology. Ecological Modelling 135, 147–186.

Guralnick, R., Hill, A., 2009. Biodiversity informatics: automated approaches for documenting global biodiversity patterns and processes. Bioinformatics 25, 421–428.

Haining, R., 2003. Spatial Data Analysis: theory and practice. Cambridge University Press.

Hannah, L., Midgley, G., Andelman, S., Araújo, M., Hughes, G., Martinez-Meyer, E., Pearson, R., Williams, P., 2007. Protected area needs in a changing climate. Front Ecol Environ 5, 131–138.

Hansen, M., DeFries, R., Townshend, J., Carroll, M., Dimiceli, C., Sohlberg, R., 2003. Global percent tree cover at a spatial resolution of 500 meters: First results of the modis vegetation continuous fields algorithm. Earth Interactions 7, 1 – 15.

Hawkins, B. A., Field, R., Cornell, H. V., Currie, D. J., Guégan, J., Kaufman, D. M., Kerr, J. T., Mittelbach, G., Oberdorff, T., O'Brien, E. M., Porter, E. E., Turner, J. R. G., 2003. Energy, water, and broad-scale geographic patterns of species richness. Ecology 84, 3105 – 3117.

Hawthorne, W., Jongkind, C., 2006. Woody Plants of Western African Forest. Kew Publishing.

Hector, A., Bagchi, R., 2007. Biodiversity and ecosystem multifunctionality. Nature 448, 188–190.

Heltshe, J., Forrester, N., 1983. Estimating species richness using the jackknife procedure. Biometrics 39, 1– 11.

Hengl, T., 2006. Finding the right pixel size. Computers & Geosciences 32 (9), 1283 – 1298.

Hengl, T., Sierdsema, H., Radovic, A., Dilo, A., 2009. Spatial prediction of species' distributions from occurrence-only records: combining point pattern analysis, enfa and

regression-kriging. Ecological Modelling 220 (24), 3499 – 3511, selected Papers on Spatially Explicit Landscape Modelling: Current practices and challenges.

Hernandez, P. A., Graham, C. H., Master, L. L., Albert, D. L., 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography 29, 773–785.

Heywood, V. H., 1993. Flowering plants of the world. University Press, New York.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., Jarvis, A., 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25, 1965 – 1978.

Hirzel, A., Helfer, V., Metral, F., 2001. Assessing habitat-suitability models with a virtual species. Ecological Modelling 145, 111–121.

Hirzel, A. H., Hausser, J., Chessel, D., Perrin, N., 2002. Ecological-niche factor analysis: How to compute habitat- suitability maps without absence data? Ecology 83, 2027–2036.

Humphries, C., Araújo, M., Williams, P., Lampinen, R., Lathi, T., Uotila, P., 1999. Plant diversity in europe: Atlas flora europaeae and worldmap. Acta botanica Fennica 162, 11–21.

Hurlbert, A. H., Jetz, W., August 2007. Species richness, hotspots, and the scale dependence of range maps in ecology and conservation. Proceedings of the National Academy of Sciences of the USA 104, 13384–13389.

Hutchinson, G. E., 1959. Homage to santa rosalia or why are there so many kinds of animals? American Naturalist 93, 145–159.

Jetz, W., Kreft, H., Ceballos, G., Mutke, J., 2008. Global associations between terrestrial producer and vertebrate consumer diversity. Proceedings of the Royal Society of London B 276, 269–278.

Jiménez, I., Distler, T., Jørgensen, P. M., 2009. Estimated plant richness pattern across northwest south america provides similar support for the species-energy and spatial heterogeneity hypotheses. Ecography 32, 433–448.

Justus, J., Sarkar, S., 2002. The principle of complementarity in the design of reserve networks to conserve biodiversity: a preliminary history. Journal of Bioscience 27, 421–435.

Kadmon, R., Farben, F., Danin, A., 2004. Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. Ecological Applications 14, 401–413.

Kearney, M., Porter, W., 2009. Mechanistic niche modelling: combining physiological and spatial data to predict species ranges. Ecology Letters 12, 334–350.

Kerr, J. T., Packer, L., 1997. Habitat heterogeneity as a determinant of mammal species richness in high-energy regions. Nature 385, 253–254.

Kier, G., Barthlott, W., 2001. Measuring and mapping endemism and species richness: a new methodological approach and its application on the flora of africa. Biodiversity and Conservation 10, 1513–1529.

Kier, G., Küper, W., Mutke, J., Rafiqpoor, M. D., Barthlott, W., 2006. African vascular plant species richness: a comparison of mapping approaches. In: Taxonomy and ecology of African plants, their conservation and sustainable use. Addis Ababa, pp. 409–425.

Kier, G., Mutke, J., Dinerstein, E., Ricketts, T. H., Küper, W., Kreft, H., Barthlott, W., 2005. Global patterns of plant diversity and floristic knowledge. Journal of Biogeography 32, 1107–1116.

Kissling, D. W., Field, R., Böhning-Gaese, K., 2008. Spatial patterns of woody plant and bird diversity: functional relationships or environmental effects? Global Ecology and Biogeography 17, 327–339.

Kreft, H., Jetz, W., March 2007. Global patterns and determinants of vascular plant diversity. PNAS 104, 5925–5930.

Kreft, H., Jetz, W., Mutke, J., Kier, G., Barthlott, W., 2008. Global diversity of island floras from a macroecological perspective. Ecology Letters 11, 116–127.

Küper, W., Sommer, J. H., Lovett, J., Barthlott, W., 2006. Deficiency in african plant distribution data-missing pieces of the puzzle. Botanical Journal of the Linnean Society 150, 355–368.

Küper, W., Sommer, J. H., Lovett, J. C., Mutke, J., Linder, H. P., Beentje, H. J., Van Rompaey, R., Chatelain, C., Sosef, M., Barthlott, W., 2004a. Africa's hotspots of biodiversity redefined. Annals of the Missouri Botanical Garden 91, 525–536.

Küper, W., Wagner, T., Barthlott, W., 2004b. Diversity patterns of plants and phytophagous beetles in sub-saharan africa. Bonner zoologische Beiträge 53, 283–289.

Lamoreux, J. F., Morrison, J. C., Ricketts, T. H., Olson, D., Dinerstein, E., McKnight, M., Shugart, H. H., 2006. Global tests of biodiversity concordance and the importance of endemism. Nature 440, 212–214.

Le Lay, G., Guisan, A., 2008. Niche-based distribution models to the rescue of rare species. In: Annual meeting of the International Congress for Conservation Biology.

Lebrun, J.-P., 1960. Sur la richesse de la flore de divers territoires africains. Bulletin de séances de l'Académie Royale des Sciences d'Outre-Mer 6, 669–690.

Legendre, P., Legendre, L., 1998. Numerical Ecology, 2nd Edition. Elsevier Science.

Linder, H. P., Lovett, J. C., Mutke, J., Barthlott, W., Jürgens, N., Rebelo, T., Küper, W., 2005. A numerical re-evaluation of the sub-saharan phytochoria of mainland africa. Biologiske Skrifter 55, 229–252.

Liu, C., Berry, P. M., Dawson, T. P., Pearson, R. G., 2005. Selecting thresholds of occurrence in the prediction of species distributions. Ecography 28, 385–393.

Loiselle, B. A., Jørgensen, P. M., Consiglio, T., Jiménez, I., Blake, J. G., Lohmann, L. G., Montiel, O. M., 2008. Predicting species distributions from herbarium collections: does climate bias in collection sampling influence model outcomes? Journal of Biogeography 35, 105–116.

Loreau, M., Naeem, S., Inchausti, P., Bengtsson, J., Grime, J. P., Hector, A., Hooper, D. U., Huston, M. A., Raffaelli, D., Schmid, B., Tilman, D., Wardle, D. A., 2001. Biodiversity and ecosystem functioning: Current knowledge and future challenges. Science 294, 804–808.

Luoto, M., Kuussaari, M., Toivonen, T., 2002. Modelling butterfly distribution based on remote sensing data. Journal of Biogeography 29, 1027–1037.

MacArthur, R. H., MacArthur, J. W., 1961. On bird species diversity. Ecology 42, 594–598.

Marsaglia, G., Tsang, W. W., Wang, J., 11 2003. Evaluating kolmogorov's distribution. Journal of Statistical Software 8 (18), 1–4.

McClean, C., Lovett, J. C., Küper, W., Hannah, L., Sommer, J. H., Barthlott, W., Termansen, M., Smith, G. F., Tokumine, S., Taplin, J., 2005. African plant diversity and climate change. Annals of the Missouri Botanical Garden 92, 139–152.

McClean, C. J., Doswald, N., Küper, W., Sommer, J. H., Barnard, P., Lovett, J. C., 2006. Potential impacts of climate change on sub-saharan african plant priority area selection. Diversity and Distributions 12, 645–655.

McKee, J. K., Sciulli, P. W., Fooce, C. D., Waite, T. A., 2004. Forecasting global biodiversity threats associated with human population growth. Biological Conservation 115 (1), 161–164.

McKinney, M. L., 2002. Urbanization, biodiversity, and conservation. BioScience 52, 883–890.

McKnight, M. W., White, P., McDonald, R. I., Lamoreux, J. F., Sechrest, W., Ridgley, R. S., Stuart, S. N., 2007. Putting beta-diversity on the map: Broad-scale congruence and coincidence in the extremes. PLOS Biology 5, 2424–2432.

McNemar, Q., 1947. Note on the sampling error of the difference between correlated proportions of percentages. Psychometrika 12, 153–157.

Millennium Ecosystem Assessment, 2005. Millennium ecosystem assessment synthesis report. Tech. rep., Millennium Ecosystem Assessment.

Mittermeier, R. A., Gil, P. R., Hoffmann, M., Pilgrim, J. D., Brooks, T. M., Mittermeier, C. G., Lamoreux, J. F., Da Fonseca, G. A. B., 2005. Hotspots revisited: Earth's biologically richest and most endangered terrestrial ecoregions. Sierra Madre.

Mutke, J., Barthlott, W., 2005. Patterns of vascular plant diversity at continental to global scales. Biologiske Skrifter 55, 521–537.

Mutke, J., Kier, G., Braun, G., Schultz, C., Barthlott, W., 2001. Patterns of african vascular plant diversity – a gis based analysis. Systematics and Geography of Plants 71, 1125–1136.

Myers, N., Mittermeier, R. A., Mittermeier, C. G., da Fonseca, G. A. B., Kent, J., 2000. Biodiversity hotspots for conservation priorities. NATURE 403, 853–858.

Nelson, B. W., Ferreira, C. A. C., d. Silva, M. F., Kawasaki, M. L., 1990. Endemism centers, refugia and botanical collection density in brazilian amazonia. NATURE 345, 714–716.

Nussbaumer, L., Gautier, L., Chatelain, C., Spichiger, R., 2005. Structure et composition floristique de la forêt classée du scio, côte d'ivoire. etude descriptive et comparative. Candollea 60, 393–502.

Olson, D. M., Dinerstein, E., Wikramanayake, E. D., Burgess, N. D., Powell, G. V., Underwood, E. C., D'Amico, J. A., Itoua, I., Strand, H. E., Morrison, J. C., Loucks, C. J., Allnut, T. F., Ricketts, T. H., Kura, Y., Lamoreux, J. F., Wettengel, W. W., Hedao, P., Kassem, K. R., 2001. Terrestrial ecoregions of the world: A new map of life on earth. BioScience 51, 933 – 938.

Orme, C. D. L., Davies, R. G., Burgess, M., Eigenbrod, F., Pickup, N., Olson, V. A., Webster, A. J., Ding, T.-S., Rasmussen, P. C., Ridgely, R. S., Stattersfield, A. J., Bennett, P. M., Blackburn, T. M., Gaston, K. J., Owens, I. P. F., 2005. Global hotspots of species richness are not congruent with endemism or threat. Nature 436, 1016–1019.

Ozenda, P., 1982. Les Végétaux dans la Biosphère. Doin Editeurs, Paris, Paris.

Palmer, M. W., 1990. The estimation of species richness by extrapolation. Ecology 71, 1195–1198.

Parnell, J. A. N., Simpson, D. A., Moat, J., Kirkup, D. W., Chantaranothai, P., Boyce, P. C., Bygrave, P., Dransield, S., Jebb, M. H. P., Macklin, J., Meade, C., Middleton, D. J., Musaya, A. M., Prajaksood, A., Pendry, C. A., Pooma, R., Suddee, S., Wilkin, P., 2003. Plant collecting spread and densities: their potential impact on biogeographical studies in thailand. Journal of Biogeography 30, 193–209.

Paton, A., 2009. Biodiversity informatics and the plant conservation baseline. Trends in Plant Science 14 (11), 629 – 637, special Issue: Plant science research in botanic gardens.

Pearce, J., Lindenmayer, D., 1998. Bioclimatic analysis to enhance reintroduction biology of the endangered helmeted honeyeater (lichenostomus melanops cassidix) in southeastern australia. Ecology 6, 238 – 243.

Pearson, D. L., Carroll, S. S., 1999. The influence of spatial scale on cross-taxon congruence patterns and prediction accuracy of species richness. Journal of Biogeography 26, 1079–1090.

Pearson, R., Dawson, T., 2003. Predicting the impacts of climate change on the distribution of species: Are bioclimate envelope models useful? Global Ecology and Biogeography 12, 361 – 371.

Pearson, R. G., Raxworthy, C. J., Nakamura, M., Townsend Peterson, A., 2007. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in madagascar. Journal of Biogeography 34, 102–117.

Phillips, S. J., Anderson, R. P., Schapire, R. E., 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling 190, 231–259.

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., Ferrier, S., 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecological Applications 19 (1), 181–197.

Phillips, S. J., Dudík, M., Schapire, R. E., 2004. A maximum entropy approach to species distribution modeling. In: Proceedings of the 21st International Conference on Machine Learning. New York, pp. 655–662.

Poorter, L., Bongers, F., Kouamé, F. N., Hawthorne, W. D., 2004a. Biodiversity of West African Forests. An Ecological Atlas of Woody Plant Species. CABI, Oxon.

Poorter, L., Bongers, F., Lemmens, R., 2004b. Biodiversity of West African Forests. An Ecological Atlas of Woody Plant Species. CABI, Ch. West African forests: introduction, pp. 5 – 14.

Prendergast, J. R., Quinn, R. M., Lawton, J. H., Eversham, B. C., Gibbons, D. W., 1993. Rare species, the coincidence of diversity hotspots and conservation strategies. Nature 365, 335–337.

R Development Core Team, 2009. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0. URL `http://www.R-project.org`

Rahbek, C., Graves, G. R., 2001. Multiscale assessment of patterns of avian species richness. Proceedings of the National Academy of Sciences of the USA 98, 4534–4539.

Rangel, T., Diniz-Filho, J., Bini, L., 2006. Towards an integrated computational tool for spatial analysis in marcoecology and biogeography. Global Ecology and Biogeography 15, 321–327.

Raxworthy, C. J., Martinez-Meyer, E., Horning, N., Nussbaum, N. A., Schneider, G. E., Ortega-Huerta, M. A., Peterson, A. T., 2004. Predicting distributions of known and unknown reptile species in madagascar. Nature, 837 – 841.

Reddy, S., Dávalos, L. M., 2003. Geographical sampling bias and its implication for conservation priorities in africa. Journal of Biogeography, 1719–1727.

Robertson, M. P., Barker, N. P., 2006. A technique for evaluating species richness maps generated from collections data. South African Journal of Science 102, 77–84.

Robertson, M. P., Peter, C. I., Villet, M. H., Ripley, B. S., 2003. Comparing models for predicting species‘ potential distributions: a case study using correlative and mechanistic predictive modelling techniques. Ecological Modelling 164, 153–167.

Rodríguez, J. P., Brotons, L., Bustamante, J., Seoane, J., 2007. The application of predictive modelling of species distribution to biodiversity conservation. Diversity and Distributions 13, 243–251.

Romo, H., García-Barros, E., Lobo, J. M., 2006. Identifying recorder-induced geographic bias in an iberian butterfly database. Ecography 29, 873–885.

Root, T. L., MacMynowski, D. P., Mastrandrea, M. D., Schneider, S. H., 2005. Human-modified temperatures induce species changes: Joint attribution. Proceedings of the National Academy of Sciences of the USA 102, 7465–7469.

SAGA Development Team, 2008. System for Automated Geoscientific Analyses (SAGA GIS). Germany.
URL `http://www.saga-gis.org/`

Sanderson, E., Jaiteh, M., Levy, M., Redford, K., Wannebo, A., Woolmer, G., 2002. The human footprint and the last of the wild. Bioscience 52, 891–904.

Sax, D. F., Gaines, S. D., 2003. Species diversity: from global decreases to local increases. Trends in Ecology and Evolution 18, 561–566.

Schabenberger, O., Gotway, C., 2005. Statistical Methods for Spatial Data Data Analysis. Chapman & Hall, London.

Schmidt, M., König, K., Müller, J., 2008. Modelling species richness and life form composition in sahelian burkina faso with remote sensing data. Journal of Arid Environments 72 (8), 1506 – 1517.

Schmidt, M., Kreft, H., Thiombiano, A., Zizka, G., 2005. Herbarium collections and field data-based plant diversity maps for burkina faso. Diversity and Distributions 11, 509–516.

Schulman, L., Toivonen, T., Ruokolainen, K., 2007. Analysing botanical collecting effort in amazonia and correcting for it in species range estimation. Journal of Biogeography 34, 1388–1399.

Scott, J., Csuti, B., Jacobi, J., Estes, J., 1987. Species richness. BioScience 37, 782–788.

Scott, J. M., Heglund, P. J., Morrison, M. L., 2002. Predicting Species Occurrences: Issues of Accuracy and Scale. Island Press, Washington, D.C.

Segurado, P., Araújo, M. B., 2004. An evaluation of methods for modelling species distributions. Journal of Biogeography 31, 1–14.

Shmida, A., Wilson, M. V., 1985. Biological determinants of species diversity. Journal of Biogeography 12, 1–20.

Slocum, T., McMaster, R., Kessler, F., Howard, H., 2005. Thematic Cartography and Geographic Visualization. Prentice Hall, Upper Saddle River NJ.

Smith, E. P., Belle, G. v., 1984. Nonparametric estimation of species richness. Biometrics 40 (1), 119–129.

Soberón, J., Jiménez, R., Golubov, J., Koleff, P., 2007. Assessing completeness of biodiversity databases at different spatial scales. Ecography 30, 152–160.

Soberón, J., Peterson, A., 2009. Monitoring biodiversity loss with primary species-occurrence data: toward national-level indicators for the 2010 target of the convention on biological diversity. Ambio 38, 29–34.

Soberón, J., Peterson, T., 2004. Biodiversity informatics: managing and applying primary biodiversity data. Phil. Trans. R. Soc. Lond. B 359, 689–698.

Soberón, J. M., Llorente, J. B., Oñate, L., 2000. The use of specimen-label databases for conservation purposes: an example using mexican papilionid and pierid butterflies. Biodiversity and Conservation, 1441–1466.

Soria-Auza, R., Kessler, M., 2008. The influence of sampling intensity on the perception of the spatial distribution of tropical diversity and endemism: a case study of ferns from bolivia. Diversity and Distributions 14, 123–130.

Stockwell, D., Peters, D., 1999. The garp modelling system: problems and solutions to automated spatial prediction. International Journal of Geographical Information Science 13, 143–158.

Stockwell, D. R. B., Peterson, A. T., 2002. Effects of sample size on accuracy of species distribution models. Ecological Modelling 148, 1–13.

Su, J. C., Debinski, D. M., Jakubauskas, M. E., Kindscher, K., 2004. Beyond species richness: Community similarity as a measure of cross-taxon congruence for coarse-filter conservation. Conservation Biology 18, 167–173.

Thiombiano, A., Schmidt, M., Kreft, H., Guinko, S., 2006. Influence du gradient climatique sur la distribution des espèces de combretaceae dau burkina faso (afrique de l'ouest). Candollea 61, 189–213.

Thuiller, W., Lavorel, S., Sykes, M. T., Araújo, M. B., 2006. Using niche-based modelling to assess the impact of climate change on tree functional diversity in europe. Diversity and Distributions 12, 49–60.

Thuiller, W., Richardson, D., Pysek, P., Midgley, G., Hughes, G., Rouget, M., 2005. Niche-based modeling as a tool for predicting the global risk of alien plant invasions. Global Change Biology 11, 2234 – 2250.

Tscharntke, T., Klein, A. M., Kruess, A., Steffan-Dewenter, I., Thies, C., 2005. Landscape perspectives on agricultural intensification and biodiversity – ecosystem service management. Ecology Letters 8, 857–874.

Usher, M., 1986. Wildlife Conservation Evaluation. Chapman & Hall, Ch. Wildlife conservation evaluation: attributes, criteria and values, pp. 3 – 44.

Veloz, S. D., 2009. Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only. Journal of Biogeography 36, 2290–2299.

Walther, B. A., Martin, J. L., 2001. Species richness estimation of bird communities: how to control for sampling effort? Ibis, 413–419.

Walther, B. A., Moore, J. L., 2005. The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. Ecography 28, 815–829.

White, F., 1983. The vegetation of Africa. Unesco, Paris.

Whittaker, R. J., Araujo, M. B., Paul, J., Ladle, R. J., Watson, J. E. M., Willis, K. J., 2005. Conservation biogeography: assessment and prospect. Diversity and Distributions 11, 3–23.

Whittaker, R. J., Willis, K. J., Field, R., 2001. Scale and species richness: towards a general, hierarchical theory of species diversity. Journal of Biogeography 28, 453–470.

Wiens, J. A., Stralberg, D., Jongsomjit, D., Howell, C. A., Snyder, M. A., 2009. Niches, models, and climate change: Assessing the assumptions and uncertainties. PNAS 106 (Supplement 2), 19729–19736.

Williams, P., 2001. Encyclopedia of Biodiversity. Academic Press, Ch. Complementarity, pp. 813–829.

Williams, P., Faith, D., Manne, L., Sechrest, W., Preston, C., 2006. Complementarity analysis: Mapping the performance of surrogates for biodiversity. Biological Conservation 128 (2), 253 – 264.

Williams, P. H., Gibbons, D., Margules, C., Rebelo, A., Humphries, C., Pressey, R. L., 1996. A comparison of richness hotspots, rarity hotspots and complementary areas for conserving diversity of british birds. Conservation Biology 10, 155–174.

Williams, P. H., Margules, C. R., Hilbert, D. W., 2002. Data requirements and data sources for biodiversity priority area selection. J. Bioscience 27, 327–338.

World Conservation Union and UNEP-World Conservation Monitoring Centre, 2007. World Database On Protected Areas. WCMC, Cambridge, UK.

Yates, D. N., Kittel, T. G. F., Cannon, R. F., 2000. Comparing the correlative holdridge model to mechanistic biogeographical models for assessing vegetation distribution response to climatic change. Climatic Change 44, 59–87.

Zaniewski, A. E., Lehmann, A., Overton, J., 2002. Predicting species spatial distributions using presence-only data: a case study of native new zealand ferns. Ecological Modelling 157, 261–280.

LIST OF FIGURES

LIST OF TABLES